Engineering Science

David Beaty

First Edition, 2012

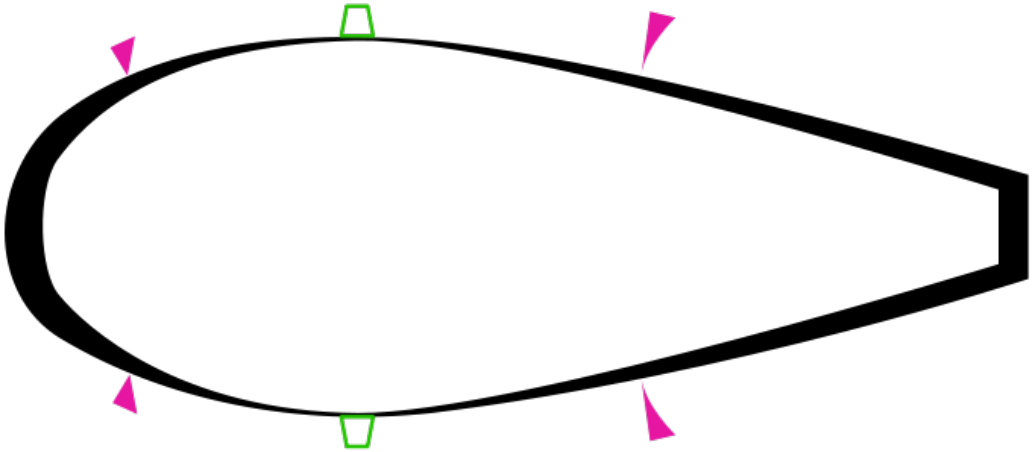
ISBN 978-81-323-4133-8

Published by:
White Word Publications
4735/22 Prakashdeep Bldg,
Ansari Road, Darya Ganj,
Delhi - 110002
Email: info@wtbooks.com

Table of Contents

Chapter 1

Fluid Dynamics



Typical aerodynamic teardrop shape, assuming a viscous medium passing from left to right, the diagram shows the pressure distribution as the thickness of the black line and shows the velocity in the boundary layer as the violet triangles. The green vortex generators prompt the transition to turbulent flow and prevent back-flow also called flow separation from the high pressure region in the back. The surface in front is as smooth as possible or even employs shark like skin, as any turbulence here will reduce the energy of the airflow. The truncation on the right, known as a Kammback, also prevents back flow from the high pressure region in the back across the spoilers to the convergent part.

In physics, **fluid dynamics** is a sub-discipline of fluid mechanics that deals with **fluid flow**—the natural science of fluids (liquids and gases) in motion. It has several subdisciplines itself, including aerodynamics (the study of air and other gases in motion) and **hydrodynamics** (the study of liquids in motion). Fluid dynamics has a wide range of applications, including calculating forces and moments on aircraft, determining the mass flow rate of petroleum through pipelines, predicting weather patterns, understanding nebulae in interstellar space and reportedly modeling fission weapon detonation. Some of its principles are even used in traffic engineering, where traffic is treated as a continuous fluid.

Fluid dynamics offers a systematic structure that underlies these practical disciplines, that embraces empirical and semi-empirical laws derived from flow measurement and used to solve practical problems. The solution to a fluid dynamics problem typically involves calculating various properties of the fluid, such as velocity, pressure, density, and temperature, as functions of space and time.

Historically, *hydrodynamics* meant something different than it does today. Before the twentieth century, hydrodynamics was synonymous with fluid dynamics. This is still reflected in names of some fluid dynamics topics, like magnetohydrodynamics and hydrodynamic stability—both also applicable in, as well as being applied to, gases.

Equations of fluid dynamics

The foundational axioms of fluid dynamics are the conservation laws, specifically, conservation of mass, conservation of linear momentum (also known as Newton's Second Law of Motion), and conservation of energy (also known as First Law of Thermodynamics). These are based on classical mechanics and are modified in quantum mechanics and general relativity. They are expressed using the Reynolds Transport Theorem.

In addition to the above, fluids are assumed to obey the *continuum assumption*. Fluids are composed of molecules that collide with one another and solid objects. However, the continuum assumption considers fluids to be continuous, rather than discrete. Consequently, properties such as density, pressure, temperature, and velocity are taken to be well-defined at infinitesimally small points, and are assumed to vary continuously from one point to another. The fact that the fluid is made up of discrete molecules is ignored.

For fluids which are sufficiently dense to be a continuum, do not contain ionized species, and have velocities small in relation to the speed of light, the momentum equations for Newtonian fluids are the Navier-Stokes equations, which is a non-linear set of differential equations that describes the flow of a fluid whose stress depends linearly on velocity gradients and pressure. The unsimplified equations do not have a general closed-form solution, so they are primarily of use in Computational Fluid Dynamics. The equations can be simplified in a number of ways, all of which make them easier to solve. Some of them allow appropriate fluid dynamics problems to be solved in closed form.

In addition to the mass, momentum, and energy conservation equations, a thermodynamical equation of state giving the pressure as a function of other thermodynamic variables for the fluid is required to completely specify the problem. An example of this would be the perfect gas equation of state:

$$p = \frac{\rho R_u T}{M}$$

where p is pressure, ρ is density, R_u is the gas constant, M is the molar mass and T is temperature.

Compressible vs incompressible flow

All fluids are compressible to some extent, that is changes in pressure or temperature will result in changes in density. However, in many situations the changes in pressure and temperature are sufficiently small that the changes in density are negligible. In this case the flow can be modeled as an incompressible flow. Otherwise the more general compressible flow equations must be used.

Mathematically, incompressibility is expressed by saying that the density ρ of a fluid parcel does not change as it moves in the flow field, i.e.,

$$\frac{\mathrm{D}\rho}{\mathrm{D}t} = 0\,,$$

where D / Dt is the substantial derivative, which is the sum of local and convective derivatives. This additional constraint simplifies the governing equations, especially in the case when the fluid has a uniform density.

For flow of gases, to determine whether to use compressible or incompressible fluid dynamics, the Mach number of the flow is to be evaluated. As a rough guide, compressible effects can be ignored at Mach numbers below approximately 0.3. For liquids, whether the incompressible assumption is valid depends on the fluid properties (specifically the critical pressure and temperature of the fluid) and the flow conditions (how close to the critical pressure the actual flow pressure becomes). Acoustic problems always require allowing compressibility, since sound waves are compression waves involving changes in pressure and density of the medium through which they propagate.

Viscous vs inviscid flow

Viscous problems are those in which fluid friction has significant effects on the fluid motion.

The Reynolds number, which is a ratio between inertial and viscous forces, can be used to evaluate whether viscous or inviscid equations are appropriate to the problem.

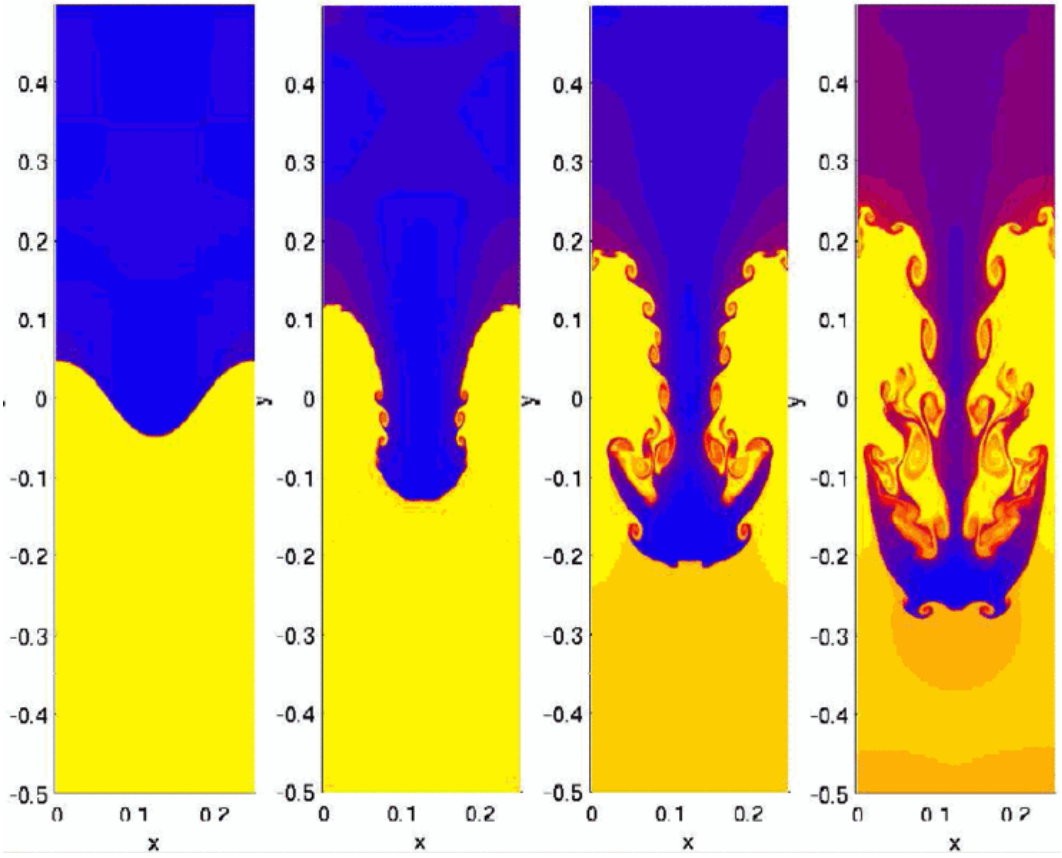
Stokes flow is flow at very low Reynolds numbers, $Re \ll 1$, such that inertial forces can be neglected compared to viscous forces.

On the contrary, high Reynolds numbers indicate that the inertial forces are more significant than the viscous (friction) forces. Therefore, we may assume the flow to be an inviscid flow, an approximation in which we neglect viscosity completely, compared to inertial terms.

This idea can work fairly well when the Reynolds number is high. However, certain problems such as those involving solid boundaries, may require that the viscosity be included. Viscosity often cannot be neglected near solid boundaries because the no-slip condition can generate a thin region of large strain rate (known as Boundary layer) which enhances the effect of even a small amount of viscosity, and thus generating vorticity. Therefore, to calculate net forces on bodies (such as wings) we should use viscous flow equations. As illustrated by d'Alembert's paradox, a body in an inviscid fluid will experience no drag force. The standard equations of inviscid flow are the Euler equations. Another often used model, especially in computational fluid dynamics, is to use the Euler equations away from the body and the boundary layer equations, which incorporates viscosity, in a region close to the body.

The Euler equations can be integrated along a streamline to get Bernoulli's equation. When the flow is everywhere irrotational and inviscid, Bernoulli's equation can be used throughout the flow field. Such flows are called potential flows.

Steady vs unsteady flow



Hydrodynamics simulation of the Rayleigh–Taylor instability

When all the time derivatives of a flow field vanish, the flow is considered to be a **steady flow**. Steady-state flow refers to the condition where the fluid properties at a point in the system do not change over time. Otherwise, flow is called unsteady. Whether a particular

flow is steady or unsteady, can depend on the chosen frame of reference. For instance, laminar flow over a sphere is steady in the frame of reference that is stationary with respect to the sphere. In a frame of reference that is stationary with respect to a background flow, the flow is unsteady.

Turbulent flows are unsteady by definition. A turbulent flow can, however, be statistically stationary. According to Pope:

The random field $U(x,t)$ is statistically stationary if all statistics are invariant under a shift in time.

This roughly means that all statistical properties are constant in time. Often, the mean field is the object of interest, and this is constant too in a statistically stationary flow.

Steady flows are often more tractable than otherwise similar unsteady flows. The governing equations of a steady problem have one dimension fewer (time) than the governing equations of the same problem without taking advantage of the steadiness of the flow field.

Laminar vs turbulent flow

Turbulence is flow characterized by recirculation, eddies, and apparent randomness. Flow in which turbulence is not exhibited is called laminar. It should be noted, however, that the presence of eddies or recirculation alone does not necessarily indicate turbulent flow—these phenomena may be present in laminar flow as well. Mathematically, turbulent flow is often represented via a Reynolds decomposition, in which the flow is broken down into the sum of an average component and a perturbation component.

It is believed that turbulent flows can be described well through the use of the Navier–Stokes equations. Direct numerical simulation (DNS), based on the Navier–Stokes equations, makes it possible to simulate turbulent flows at moderate Reynolds numbers. Restrictions depend on the power of the computer used and the efficiency of the solution algorithm. The results of DNS have been found to agree well with experimental data for some flows.

Most flows of interest have Reynolds numbers much too high for DNS to be a viable option, given the state of computational power for the next few decades. Any flight vehicle large enough to carry a human ($L > 3$ m), moving faster than 72 km/h (20 m/s) is well beyond the limit of DNS simulation ($Re = 4$ million). Transport aircraft wings (such as on an Airbus A300 or Boeing 747) have Reynolds numbers of 40 million (based on the wing chord). In order to solve these real-life flow problems, turbulence models will be a necessity for the foreseeable future. Reynolds-averaged Navier–Stokes equations (RANS) combined with turbulence modeling provides a model of the effects of the turbulent flow. Such a modeling mainly provides the additional momentum transfer by

the Reynolds stresses, although the turbulence also enhances the heat and mass transfer. Another promising methodology is large eddy simulation (LES), especially in the guise of detached eddy simulation (DES)—which is a combination of RANS turbulence modeling and large eddy simulation.

Newtonian vs non-Newtonian fluids

Sir Isaac Newton showed how stress and the rate of strain are very close to linearly related for many familiar fluids, such as water and air. These Newtonian fluids are modeled by a coefficient called viscosity, which depends on the specific fluid.

However, some of the other materials, such as emulsions and slurries and some visco-elastic materials (e.g. blood, some polymers), have more complicated *non-Newtonian* stress-strain behaviours. These materials include *sticky liquids* such as latex, honey, and lubricants which are studied in the sub-discipline of rheology.

Subsonic vs transonic, supersonic and hypersonic flows

While many terrestrial flows (e.g. flow of water through a pipe) occur at low mach numbers, many flows of practical interest (e.g. in aerodynamics) occur at high fractions of the Mach Number M=1 or in excess of it (supersonic flows). New phenomena occur at these Mach number regimes (e.g. shock waves for supersonic flow, transonic instability in a regime of flows with M nearly equal to 1, non-equilibrium chemical behavior due to ionization in hypersonic flows) and it is necessary to treat each of these flow regimes separately.

Magnetohydrodynamics

Magnetohydrodynamics is the multi-disciplinary study of the flow of electrically conducting fluids in electromagnetic fields. Examples of such fluids include plasmas, liquid metals, and salt water. The fluid flow equations are solved simultaneously with Maxwell's equations of electromagnetism.

Other approximations

There are a large number of other possible approximations to fluid dynamic problems. Some of the more commonly used are listed below.

- The **Boussinesq approximation** neglects variations in density except to calculate buoyancy forces. It is often used in free convection problems where density changes are small.
- **Lubrication theory** and **Hele-Shaw flow** exploits the large aspect ratio of the domain to show that certain terms in the equations are small and so can be neglected.
- **Slender-body theory** is a methodology used in Stokes flow problems to estimate the force on, or flow field around, a long slender object in a viscous fluid.

- The **shallow-water equations** can be used to describe a layer of relatively inviscid fluid with a free surface, in which surface gradients are small.
- The **Boussinesq equations** are applicable to surface waves on thicker layers of fluid and with steeper surface slopes.
- **Darcy's law** is used for flow in porous media, and works with variables averaged over several pore-widths.
- In rotating systems, the **quasi-geostrophic approximation** assumes an almost perfect balance between pressure gradients and the Coriolis force. It is useful in the study of atmospheric dynamics.

Terminology in fluid dynamics

The concept of pressure is central to the study of both fluid statics and fluid dynamics. A pressure can be identified for every point in a body of fluid, regardless of whether the fluid is in motion or not. Pressure can be measured using an aneroid, Bourdon tube, mercury column, or various other methods.

Some of the terminology that is necessary in the study of fluid dynamics is not found in other similar areas of study. In particular, some of the terminology used in fluid dynamics is not used in fluid statics.

Terminology in incompressible fluid dynamics

The concepts of total pressure and dynamic pressure arise from Bernoulli's equation and are significant in the study of all fluid flows. (These two pressures are not pressures in the usual sense—they cannot be measured using an aneroid, Bourdon tube or mercury column.) To avoid potential ambiguity when referring to pressure in fluid dynamics, many authors use the term static pressure to distinguish it from total pressure and dynamic pressure. Static pressure is identical to pressure and can be identified for every point in a fluid flow field.

In *Aerodynamics*, L.J. Clancy writes: *To distinguish it from the total and dynamic pressures, the actual pressure of the fluid, which is associated not with its motion but with its state, is often referred to as the static pressure, but where the term pressure alone is used it refers to this static pressure.*

A point in a fluid flow where the flow has come to rest (i.e. speed is equal to zero adjacent to some solid body immersed in the fluid flow) is of special significance. It is of such importance that it is given a special name—a stagnation point. The static pressure at the stagnation point is of special significance and is given its own name—stagnation pressure. In incompressible flows, the stagnation pressure at a stagnation point is equal to the total pressure throughout the flow field.

Terminology in compressible fluid dynamics

In a compressible fluid, such as air, the temperature and density are essential when determining the state of the fluid. In addition to the concept of total pressure (also known as stagnation pressure), the concepts of total (or stagnation) temperature and total (or stagnation) density are also essential in any study of compressible fluid flows. To avoid potential ambiguity when referring to temperature and density, many authors use the terms static temperature and static density. Static temperature is identical to temperature; and static density is identical to density; and both can be identified for every point in a fluid flow field.

The temperature and density at a stagnation point are called stagnation temperature and stagnation density.

A similar approach is also taken with the thermodynamic properties of compressible fluids. Many authors use the terms total (or stagnation) enthalpy and total (or stagnation) entropy. The terms static enthalpy and static entropy appear to be less common, but where they are used they mean nothing more than enthalpy and entropy respectively, and the prefix "static" is being used to avoid ambiguity with their 'total' or 'stagnation' counterparts. Because the 'total' flow conditions are defined by isentropically bringing the fluid to rest, the total (or stagnation) entropy is by definition always equal to the "static" entropy.

Chapter 2

Solid Mechanics and Solid State (Electronics)

Solid mechanics

Solid mechanics is the branch of mechanics, physics, and mathematics that concerns the behavior of solid matter under external actions (e.g., external forces, temperature changes, applied displacements, etc.). It is part of a broader study known as continuum mechanics. One of the most common practical applications of solid mechanics is the Euler-Bernoulli beam equation. Solid mechanics extensively uses tensors to describe stresses, strains, and the relationship between them.

Relationship to continuum mechanics

As shown in the following table, solid mechanics inhabits a central place within continuum mechanics. The field of rheology presents an overlap between solid and fluid mechanics.

<table>
<tr><td rowspan="4">Continuum mechanics
The study of the physics of continuous materials</td><td rowspan="2">Solid mechanics
The study of the physics of continuous materials with a defined rest shape.</td><td colspan="2">Elasticity
Describes materials that return to their rest shape after an applied stress.</td></tr>
<tr><td>Plasticity
Describes materials that permanently deform after a sufficient applied stress.</td><td rowspan="2">Rheology
The study of materials with both solid and fluid characteristics.</td></tr>
<tr><td rowspan="2">Fluid mechanics
The study of the physics of continuous materials which take the shape of their container.</td><td>Non-Newtonian fluids</td></tr>
<tr><td colspan="2">Newtonian fluids</td></tr>
</table>

Response models

A material has a rest shape and its shape departs away from the rest shape due to stress. The amount of departure from rest shape is called deformation, the proportion of deformation to original size is called strain. If the applied stress is sufficiently low (or the imposed strain is small enough), almost all solid materials behave in such a way that the strain is directly proportional to the stress; the coefficient of the proportion is called the modulus of elasticity or Young's modulus. This region of deformation is known as the linearly elastic region.

It is most common for analysts in solid mechanics to use linear material models, due to ease of computation. However, real materials often exhibit non-linear behavior. As new materials are used and old ones are pushed to their limits, non-linear material models are becoming more common.

There are three models that describe how a solid responds to an applied stress:

1. Elastically – When an applied stress is removed, the material returns to its undeformed state. Linearly elastic materials, those that deform proportionally to the applied load, can be described by the linear elasticity equations such as Hooke's law.
2. Viscoelastically – These are materials that behave elastically, but also have damping: when the stress is applied and removed, work has to be done against the damping effects and is converted in heat within the material resulting in a hysteresis loop in the stress–strain curve. This implies that the material response has time-dependence.
3. Plastically – Materials that behave elastically generally do so when the applied stress is less than a yield value. When the stress is greater than the yield stress, the material behaves plastically and does not return to its previous state. That is, deformation that occurs after yield is permanent.

Solid state (electronics)

Solid-state electronics are those circuits or devices built entirely from solid materials and in which the electrons, or other charge carriers, are confined entirely within the solid material. The term is often used to contrast with the earlier technologies of vacuum and gas-discharge tube devices and it is also conventional to exclude electro-mechanical devices (relays, switches, hard drives and other devices with moving parts) from the term solid state. While solid-state can include crystalline, polycrystalline and amorphous solids and refer to electrical conductors, insulators and semiconductors, the building material is most often crystalline semiconductor. Common solid-state devices include transistors, microprocessor chips, and DRAM. DRAM devices are used in computers, flash drives

and more recently, solid state drives to replace mechanically rotating magnetic disc hard drives. A considerable amount of electromagnetic and quantum-mechanical action takes place within the device. The expression became prevalent in the 1950s and the 1960s, during the transition from vacuum tube technology to semiconductor diodes and transistors. More recently, the integrated circuit (IC), the light-emitting diode (LED), and the liquid-crystal display (LCD) have evolved as further examples of solid-state devices.

In a solid-state component, the current is confined to solid elements and compounds engineered specifically to switch and amplify it. Current flow can be understood in two forms: as negatively-charged electrons, and as positively-charged electron deficiencies called electron holes or just "holes". In some semiconductors, the current consists mostly of electrons; in other semiconductors, it consists mostly of "holes". Both the electron and the hole are called charge carriers.

For data storage, solid-state devices are much faster and more reliable but are usually more expensive. Although solid-state costs continually drop, disks, tapes, and optical disks also continue to improve their cost/performance ratio.

The first solid-state device was the "cat's whisker" detector, first used in 1930s radio receivers. A whisker-like wire was moved around on a solid crystal (such as a germanium crystal) in order to detect a radio signal. The solid-state device came into its own with the invention of the transistor in 1947.

Chapter 3

Operations Research

Operations research (also referred to as **decision science**, or **management science**) is an interdisciplinary mathematical science that focuses on the effective *use* of technology by organizations. In contrast, many other science & engineering disciplines focus on technology giving secondary considerations to its use.

Employing techniques from other mathematical sciences — such as mathematical modeling, statistical analysis, and mathematical optimization — operations research arrives at optimal or near-optimal solutions to complex decision-making problems. Because of its emphasis on human-technology interaction and because of its focus on practical applications, operations research has overlap with other disciplines, notably industrial engineering and management science, and draws on psychology and organization science. Operations Research is often concerned with determining the maximum (of profit, performance, or yield) or minimum (of loss, risk, or cost) of some real-world objective. Originating in military efforts before World War II, its techniques have grown to concern problems in a variety of industries.

Overview

Operational research encompasses a wide range of problem-solving techniques and methods applied in the pursuit of improved decision-making and efficiency. Some of the tools used by operational researchers are statistics, optimization, probability theory, queuing theory, game theory, graph theory, decision analysis, mathematical modeling and simulation. Because of the computational nature of these fields, OR also has strong ties to computer science and analytics. Operational researchers faced with a new problem must determine which of these techniques are most appropriate given the nature of the system, the goals for improvement, and constraints on time and computing power.

Work in operational research and management science may be characterized as one of three categories:

- Fundamental or foundational work takes place in three mathematical disciplines: probability, optimization, and dynamical systems theory.

- Modeling work is concerned with the construction of models, analyzing them mathematically, implementing them on computers, solving them using software tools, and assessing their effectiveness with data. This level is mainly instrumental, and driven mainly by statistics and econometrics.
- Application work in operational research, like other engineering and economics' disciplines, attempts to use models to make a practical impact on real-world problems.

The major subdisciplines in modern operational research, as identified by the journal *Operations Research*, are:

- Computing and information technologies
- Decision analysis
- Environment, energy, and natural resources
- Financial engineering
- Manufacturing, service sciences, and supply chain management
- Marketing Engineering
- Policy modeling and public sector work
- Revenue management
- Simulation
- Stochastic models
- Transportation

History

As a formal discipline, operational research originated in the efforts of military planners during World War II. In the decades after the war, the techniques began to be applied more widely to problems in business, industry and society. Since that time, operational research has expanded into a field widely used in industries ranging from petrochemicals to airlines, finance, logistics, and government, moving to a focus on the development of mathematical models that can be used to analyze and optimize complex systems, and has become an area of active academic and industrial research.

Historical origins

In the World War II era, operational research was defined as "a scientific method of providing executive departments with a quantitative basis for decisions regarding the operations under their control." Other names for it included operational analysis (UK Ministry of Defence from 1962) and quantitative management.

Prior to the formal start of the field, early work in operational research was carried out by individuals such as Charles Babbage. His research into the cost of transportation and sorting of mail led to England's universal "Penny Post" in 1840, and studies into the dynamical behaviour of railway vehicles in defence of the GWR's broad gauge. Percy Bridgman brought operational research to bear on problems in physics in the 1920s and

would later attempt to extend these to the social sciences. The modern field of operational research arose during World War II.

Modern operational research originated at the Bawdsey Research Station in the UK in 1937 and was the result of an initiative of the station's superintendent, A. P. Rowe. Rowe conceived the idea as a means to analyse and improve the working of the UK's early warning radar system, Chain Home (CH). Initially, he analyzed the operating of the radar equipment and its communication networks, expanding later to include the operating personnel's behaviour. This revealed unappreciated limitations of the CH network and allowed remedial action to be taken.

Scientists in the United Kingdom including Patrick Blackett later Lord Blackett OM PRS, Cecil Gordon, C. H. Waddington, Owen Wansbrough-Jones, Frank Yates, Jacob Bronowski and Freeman Dyson, and in the United States with George Dantzig looked for ways to make better decisions in such areas as logistics and training schedules. After the war it began to be applied to similar problems in industry.

Second World War

Patrick Blackett

During the Second World War close to 1,000 men and women in Britain were engaged in operational research. About 200 operational research scientists worked for the British Army.

Patrick Blackett worked for several different organizations during the war. Early in the war while working for the Royal Aircraft Establishment (RAE) he set up a team known as the "Circus" which helped to reduce the number of anti-aircraft artillery rounds needed to shoot down an enemy aircraft from an average of over 20,000 at the start of the Battle of Britain to 4,000 in 1941.

In 1941 Blackett moved from the RAE to the Navy, first to the Royal Navy's Coastal Command, in 1941 and then early in 1942 to the Admiralty. Blackett's team at Coastal Command's Operational Research Section (CC-ORS) included two future Nobel prize winners and many other people who went on to be preeminent in their fields. They undertook a number of crucial analyses that aided the war effort. Britain introduced the

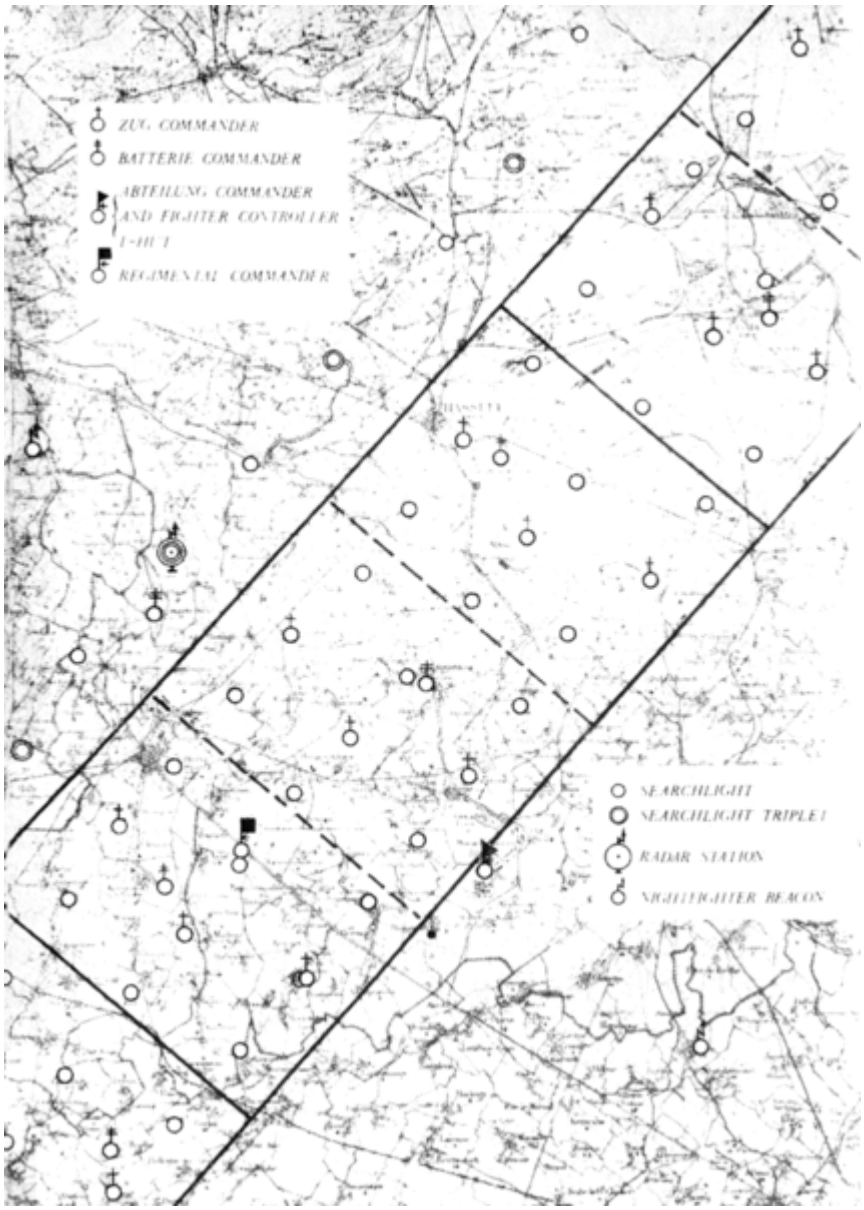
convoy system to reduce shipping losses, but while the principle of using warships to accompany merchant ships was generally accepted, it was unclear whether it was better for convoys to be small or large. Convoys travel at the speed of the slowest member, so small convoys can travel faster. It was also argued that small convoys would be harder for German U-boats to detect. On the other hand, large convoys could deploy more warships against an attacker. Blackett's staff showed that the losses suffered by convoys depended largely on the number of escort vessels present, rather than on the overall size of the convoy. Their conclusion, therefore, was that a few large convoys are more defensible than many small ones.

While performing an analysis of the methods used by RAF Coastal Command to hunt and destroy submarines, one of the analysts asked what colour the aircraft were. As most of them were from Bomber Command they were painted black for nighttime operations. At the suggestion of CC-ORS a test was run to see if that was the best colour to camouflage the aircraft for daytime operations in the grey North Atlantic skies. Tests showed that aircraft painted white were on average not spotted until they were 20% closer than those painted black. This change indicated that 30% more submarines would be attacked and sunk for the same number of sightings.

Other work by the CC-ORS indicated that on average if the trigger depth of aerial delivered depth charges (DCs) was changed from 100 feet to 25 feet, the kill ratios would go up. The reason was that if a U-boat saw an aircraft only shortly before it arrived over the target then at 100 feet the charges would do no damage (because the U-boat wouldn't have time to descend as far as 100 feet), and if it saw the aircraft a long way from the target it had time to alter course under water so the chances of it being within the 20 feet kill zone of the charges was small. It was more efficient to attack those submarines close to the surface when these targets' locations were better known than to attempt their destruction at greater depths when their positions could only be guessed. Before the change of settings from 100 feet to 25 feet, 1% of submerged U-boats were sunk and 14% damaged. After the change, 7% were sunk and 11% damaged. (If submarines were caught on the surface, even if attacked shortly after submerging, the numbers rose to 11% sunk and 15% damaged). Blackett observed "there can be few cases where such a great operational gain had been obtained by such a small and simple change of tactics".

Bomber Command's Operational Research Section (BC-ORS), analysed a report of a survey carried out by RAF Bomber Command. For the survey, Bomber Command inspected all bombers returning from bombing raids over Germany over a particular period. All damage inflicted by German air defences was noted and the recommendation was given that armour be added in the most heavily damaged areas. Their suggestion to remove some of the crew so that an aircraft loss would result in fewer personnel loss was rejected by RAF command. Blackett's team instead made the surprising and counter-intuitive recommendation that the armour be placed in the areas which were completely untouched by damage in the bombers which returned. They reasoned that the survey was biased, since it only included aircraft that returned to Britain. The untouched areas of returning aircraft were probably vital areas, which, if hit, would result in the loss of the aircraft.

(Info to help verify some/most of the above paragraph is found in "Dirty Little Secrets of the Twentieth Century" authored by James F. Dunnigan on pages 215-217



Map of *Kammhuber Line*

When Germany organised its air defences into the Kammhuber Line, it was realised that if the RAF bombers were to fly in a bomber stream they could overwhelm the night fighters who flew in individual cells directed to their targets by ground controllers. It was then a matter of calculating the statistical loss from collisions against the statistical loss from night fighters to calculate how close the bombers should fly to minimise RAF losses.

The "exchange rate" ratio of output to input was a characteristic feature of operational research. By comparing the number of flying hours put in by Allied aircraft to the number of U-boat sightings in a given area, it was possible to redistribute aircraft to more productive patrol areas. Comparison of exchange rates established "effectiveness ratios" useful in planning. The ratio of 60 mines laid per ship sunk was common to several campaigns: German mines in British ports, British mines on German routes, and United States mines in Japanese routes.

Operational research doubled the on-target bomb rate of B-29s bombing Japan from the Marianas Islands by increasing the training ratio from 4 to 10 percent of flying hours; revealed that wolf-packs of three United States submarines were the most effective number to enable all members of the pack to engage targets discovered on their individual patrol stations; revealed that glossy enamel paint was more effective camouflage for night fighters than traditional dull camouflage paint finish, and the smooth paint finish increased airspeed by reducing skin friction.

On land, the operational research sections of the Army Operational Research Group (AORG) of the Ministry of Supply (MoS) were landed in Normandy in 1944, and they followed British forces in the advance across Europe. They analysed, among other topics, the effectiveness of artillery, aerial bombing, and anti-tank shooting.

After World War II

With expanded techniques and growing awareness of the field at the close of the war, operational research was no longer limited to only operational, but was extended to encompass equipment procurement, training, logistics and infrastructure.

Academic Denis Bouyssou describes the historical development of operational research from the 1940s to the 1970s as follows. "The historical development of Operational Research (OR) is traditionally seen as the succession of several phases: the 'heroic times' of the Second World War, the 'Golden Age' between the fifties and the sixties during which major theoretical achievements were accompanied by a widespread diffusion of OR techniques in private and public organisations, a 'crisis' followed by a 'decline' starting with the late sixties, a phase during which OR groups in firms progressively disappeared while academia became less and less concerned with the applicability of the techniques developed".

Individuals such as Stafford Beer and George Dantzig pioneered early academic efforts in operational research.

Problems addressed with operational research

- critical path analysis or project planning: identifying those processes in a complex project which affect the overall duration of the project
- floorplanning: designing the layout of equipment in a factory or components on a computer chip to reduce manufacturing time (therefore reducing cost)

- network optimization: for instance, setup of telecommunications networks to maintain quality of service during outages
- allocation problems
- Bayesian search theory : looking for a target
- optimal search
- routing, such as determining the routes of buses so that as few buses are needed as possible
- supply chain management: managing the flow of raw materials and products based on uncertain demand for the finished products
- efficient messaging and customer response tactics
- automation: automating or integrating robotic systems in human-driven operations processes
- globalization: globalizing operations processes in order to take advantage of cheaper materials, labor, land or other productivity inputs
- transportation: managing freight transportation and delivery systems (Examples: LTL Shipping, intermodal freight transport)
- scheduling:
    - personnel staffing
    - manufacturing steps
    - project tasks
    - network data traffic: these are known as queueing models or queueing systems.
    - sports events and their television coverage
- blending of raw materials in oil refineries
- determining optimal prices, in many retail and B2B settings, within the disciplines of pricing science

Operational research is also used extensively in government where evidence-based policy is used.

Management science

In 1967 Stafford Beer characterized the field of management science as "the business use of operations research". However, in modern times the term management science may also be used to refer to the separate fields of organizational studies or corporate strategy. Like operational research itself, management science (MS), is an interdisciplinary branch of applied mathematics devoted to optimal decision planning, with strong links with economics, business, engineering, and other sciences. It uses various scientific research-based principles, strategies, and analytical methods including mathematical modeling, statistics and numerical algorithms to improve an organization's ability to enact rational and meaningful management decisions by arriving at optimal or near optimal solutions to complex decision problems. In short, management sciences help businesses to achieve their goals using the scientific methods of operational research.

The management scientist's mandate is to use rational, systematic, science-based techniques to inform and improve decisions of all kinds. Of course, the techniques of

management science are not restricted to business applications but may be applied to military, medical, public administration, charitable groups, political groups or community groups.

Management science is concerned with developing and applying models and concepts that may prove useful in helping to illuminate management issues and solve managerial problems, as well as designing and developing new and better models of organizational excellence.

The application of these models within the corporate sector became known as Management science.

Techniques

Some of the fields that have considerable overlap with Management Science include:

- Data mining
- Decision analysis
- Engineering
- Forecasting
- Game theory
- Industrial engineering
- Logistics
- Mathematical modeling
- Optimization
- Probability and statistics
- Project management
- Simulation
- Social network/Transportation forecasting models
- Supply chain management
- Financial engineering

Applications of management science

Applications of management science are abundant in industry as airlines, manufacturing companies, service organizations, military branches, and in government. The range of problems and issues to which management science has contributed insights and solutions is vast. It includes:

- scheduling airlines, including both planes and crew,
- deciding the appropriate place to site new facilities such as a warehouse, factory or fire station,
- managing the flow of water from reservoirs,
- identifying possible future development paths for parts of the telecommunications industry,
- establishing the information needs and appropriate systems to supply them within the health service, and
- identifying and understanding the strategies adopted by companies for their information systems

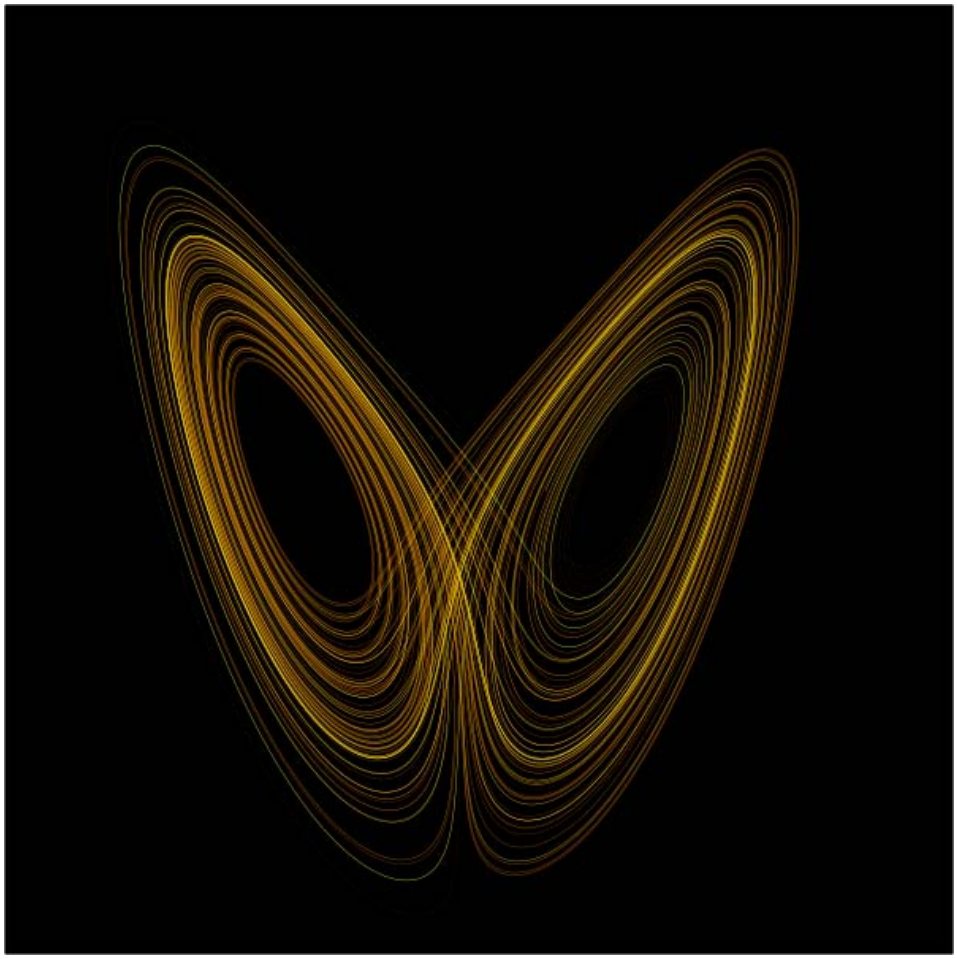
Management science is also concerned with so-called ”soft-operational analysis”, which concerns methods for strategic planning, strategic decision support, and Problem

Structuring Methods (PSM). In dealing with these sorts of challenges mathematical modeling and simulation are not appropriate or will not suffice. Therefore, during the past 30 years, a number of non-quantified modeling methods have been developed. These include:

- stakeholder based approaches including metagame analysis and drama theory
- morphological analysis and various forms of influence diagrams.
- approaches using cognitive mapping
- the Strategic Choice Approach
- robustness analysis

Chapter 4

Dynamical System



The Lorenz attractor is an example of a non-linear dynamical system. Studying this system helped give rise to Chaos theory.

A **dynamical system** is a concept in mathematics where a fixed rule describes the time dependence of a point in a geometrical space. Examples include the mathematical models that describe the swinging of a clock pendulum, the flow of water in a pipe, and the number of fish each spring in a lake.

At any given time a dynamical system has a *state* given by a set of real numbers (a vector) which can be represented by a point in an appropriate *state space* (a geometrical manifold). Small changes in the state of the system correspond to small changes in the numbers. The *evolution rule* of the dynamical system is a fixed rule that describes what future states follow from the current state. The rule is deterministic; in other words, for a given time interval only one future state follows from the current state.

Overview

The concept of a dynamical system has its origins in Newtonian mechanics. There, as in other natural sciences and engineering disciplines, the evolution rule of dynamical systems is given implicitly by a relation that gives the state of the system only a short time into the future. (The relation is either a differential equation, difference equation or other time scale.) To determine the state for all future times requires iterating the relation many times—each advancing time a small step. The iteration procedure is referred to as *solving the system* or *integrating the system*. Once the system can be solved, given an initial point it is possible to determine all its future points, a collection known as a *trajectory* or *orbit*.

Before the advent of fast computing machines, solving a dynamical system required sophisticated mathematical techniques and could be accomplished only for a small class of dynamical systems. Numerical methods implemented on electronic computing machines have simplified the task of determining the orbits of a dynamical system.

For simple dynamical systems, knowing the trajectory is often sufficient, but most dynamical systems are too complicated to be understood in terms of individual trajectories. The difficulties arise because:

- The systems studied may only be known approximately—the parameters of the system may not be known precisely or terms may be missing from the equations. The approximations used bring into question the validity or relevance of numerical solutions. To address these questions several notions of stability have been introduced in the study of dynamical systems, such as Lyapunov stability or structural stability. The stability of the dynamical system implies that there is a class of models or initial conditions for which the trajectories would be equivalent. The operation for comparing orbits to establish their equivalence changes with the different notions of stability.
- The type of trajectory may be more important than one particular trajectory. Some trajectories may be periodic, whereas others may wander through many different states of the system. Applications often require enumerating these classes or maintaining the system within one class. Classifying all possible trajectories has

led to the qualitative study of dynamical systems, that is, properties that do not change under coordinate changes. Linear dynamical systems and systems that have two numbers describing a state are examples of dynamical systems where the possible classes of orbits are understood.

- The behavior of trajectories as a function of a parameter may be what is needed for an application. As a parameter is varied, the dynamical systems may have bifurcation points where the qualitative behavior of the dynamical system changes. For example, it may go from having only periodic motions to apparently erratic behavior, as in the transition to turbulence of a fluid.
- The trajectories of the system may appear erratic, as if random. In these cases it may be necessary to compute averages using one very long trajectory or many different trajectories. The averages are well defined for ergodic systems and a more detailed understanding has been worked out for hyperbolic systems. Understanding the probabilistic aspects of dynamical systems has helped establish the foundations of statistical mechanics and of chaos.

It was in the work of Poincaré that these dynamical systems themes developed.

Basic definitions

A dynamical system is a manifold *M* called the phase (or state) space endowed with a family of smooth evolution functions Φ^t that for any element of $t \in T$, the time, map a point of the phase space back into the phase space. The notion of smoothness changes with applications and the type of manifold. There are several choices for the set *T*. When *T* is taken to be the reals, the dynamical system is called a *flow*; and if *T* is restricted to the non-negative reals, then the dynamical system is a *semi-flow*. When *T* is taken to be the integers, it is a *cascade* or a *map*; and the restriction to the non-negative integers is a *semi-cascade*.

Examples

The evolution function Φ^t is often the solution of a *differential equation of motion*

$$\dot{x} = v(x)\,.$$

The equation gives the time derivative, represented by the dot, of a trajectory *x(t)* on the phase space starting at some point x_0. The *vector field v(x)* is a smooth function that at every point of the phase space *M* provides the velocity vector of the dynamical system at that point. (These vectors are not vectors in the phase space *M*, but in the tangent space T_xM of the point *x*.) Given a smooth Φ^t, an autonomous vector field can be derived from it.

There is no need for higher order derivatives in the equation, nor for time dependence in *v(x)* because these can be eliminated by considering systems of higher dimensions. Other types of differential equations can be used to define the evolution rule:

$$G(x, \dot{x}) = 0$$

is an example of an equation that arises from the modeling of mechanical systems with complicated constraints.

The differential equations determining the evolution function Φ^t are often ordinary differential equations: in this case the phase space M is a finite dimensional manifold. Many of the concepts in dynamical systems can be extended to infinite-dimensional manifolds—those that are locally Banach spaces—in which case the differential equations are partial differential equations. In the late 20th century the dynamical system perspective to partial differential equations started gaining popularity.

Further examples

- Logistic map
- Dyadic transformation
- Tent map
- Double pendulum
- Arnold's cat map
- Horseshoe map
- Baker's map is an example of a chaotic piecewise linear map
- Billiards and outer billiards
- Hénon map
- Lorenz system
- Circle map
- Rössler map
- List of chaotic maps
- Swinging Atwood's machine
- Quadratic map simulation system
- Bouncing ball dynamics

Linear dynamical systems

Linear dynamical systems can be solved in terms of simple functions and the behavior of all orbits classified. In a linear system the phase space is the N-dimensional Euclidean space, so any point in phase space can be represented by a vector with N numbers. The analysis of linear systems is possible because they satisfy a superposition principle: if *u(t)* and *w(t)* satisfy the differential equation for the vector field (but not necessarily the initial condition), then so will *u(t) + w(t)*.

Flows

For a flow, the vector field $\Phi(x)$ is a linear function of the position in the phase space, that is,

$$\phi(x) = Ax + b\,,$$

with A a matrix, b a vector of numbers and x the position vector. The solution to this system can be found by using the superposition principle (linearity). The case $b \neq 0$ with $A = 0$ is just a straight line in the direction of b:

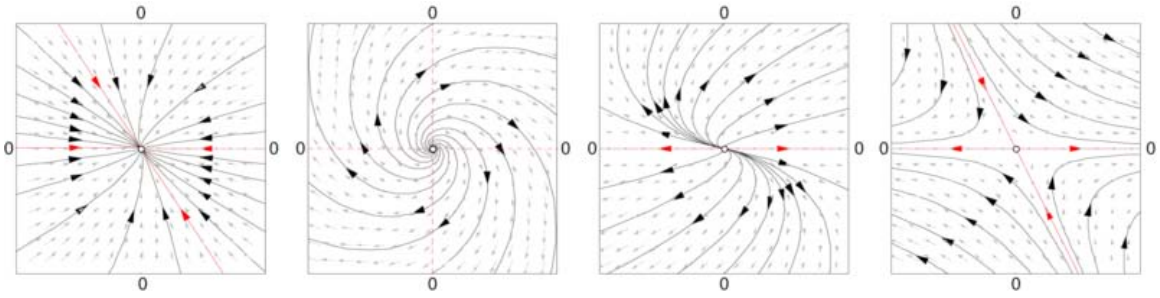
$$\Phi^t(x_1) = x_1 + bt\ .$$

When b is zero and $A \neq 0$ the origin is an equilibrium (or singular) point of the flow, that is, if $x_0 = 0$, then the orbit remains there. For other initial conditions, the equation of motion is given by the exponential of a matrix: for an initial point x_0,

$$\Phi^t(x_0) = e^{tA}x_0\ .$$

When $b = 0$, the eigenvalues of A determine the structure of the phase space. From the eigenvalues and the eigenvectors of A it is possible to determine if an initial point will converge or diverge to the equilibrium point at the origin.

The distance between two different initial conditions in the case $A \neq 0$ will change exponentially in most cases, either converging exponentially fast towards a point, or diverging exponentially fast. Linear systems display sensitive dependence on initial conditions in the case of divergence. For nonlinear systems this is one of the (necessary but not sufficient) conditions for chaotic behavior.



Linear vector fields and a few trajectories.

Maps

A discrete-time, affine dynamical system has the form

$$x_{n+1} = Ax_n + b\,,$$

with A a matrix and b a vector. As in the continuous case, the change of coordinates $x \rightarrow x + (I - A)^{-1}b$ removes the term b from the equation. In the new coordinate system, the origin is a fixed point of the map and the solutions are of the linear system $A^n x_0$. The solutions for the map are no longer curves, but points that hop in the phase space. The orbits are organized in curves, or fibers, which are collections of points that map into themselves under the action of the map.

As in the continuous case, the eigenvalues and eigenvectors of A determine the structure of phase space. For example, if u_1 is an eigenvector of A, with a real eigenvalue smaller than one, then the straight lines given by the points along $\alpha\, u_1$, with $\alpha \in \mathbf{R}$, is an invariant curve of the map. Points in this straight line run into the fixed point.

There are also many other discrete dynamical systems.

Local dynamics

The qualitative properties of dynamical systems do not change under a smooth change of coordinates (this is sometimes taken as a definition of qualitative): a *singular point* of the vector field (a point where $v(x) = 0$) will remain a singular point under smooth transformations; a *periodic orbit* is a loop in phase space and smooth deformations of the phase space cannot alter it being a loop. It is in the neighborhood of singular points and periodic orbits that the structure of a phase space of a dynamical system can be well understood. In the qualitative study of dynamical systems, the approach is to show that there is a change of coordinates (usually unspecified, but computable) that makes the dynamical system as simple as possible.

Rectification

A flow in most small patches of the phase space can be made very simple. If y is a point where the vector field $v(y) \neq 0$, then there is a change of coordinates for a region around y where the vector field becomes a series of parallel vectors of the same magnitude. This is known as the rectification theorem.

The rectification theorem says that away from singular points the dynamics of a point in a small patch is a straight line. The patch can sometimes be enlarged by stitching several patches together, and when this works out in the whole phase space M the dynamical system is *integrable*. In most cases the patch cannot be extended to the entire phase space. There may be singular points in the vector field (where $v(x) = 0$); or the patches may become smaller and smaller as some point is approached. The more subtle reason is a global constraint, where the trajectory starts out in a patch, and after visiting a series of other patches comes back to the original one. If the next time the orbit loops around phase space in a different way, then it is impossible to rectify the vector field in the whole series of patches.

Near periodic orbits

In general, in the neighborhood of a periodic orbit the rectification theorem cannot be used. Poincaré developed an approach that transforms the analysis near a periodic orbit to the analysis of a map. Pick a point x_0 in the orbit γ and consider the points in phase space in that neighborhood that are perpendicular to $v(x_0)$. These points are a Poincaré section $S(\gamma, x_0)$, of the orbit. The flow now defines a map, the Poincaré map $F : S \to S$, for points starting in S and returning to S. Not all these points will take the same amount of time to come back, but the times will be close to the time it takes x_0.

The intersection of the periodic orbit with the Poincaré section is a fixed point of the Poincaré map F. By a translation, the point can be assumed to be at $x = 0$. The Taylor series of the map is $F(x) = J \cdot x + O(x^2)$, so a change of coordinates h can only be expected to simplify F to its linear part

$$h^{-1} \circ F \circ h(x) = J \cdot x \, .$$

This is known as the conjugation equation. Finding conditions for this equation to hold has been one of the major tasks of research in dynamical systems. Poincaré first approached it assuming all functions to be analytic and in the process discovered the non-resonant condition. If $\lambda_1,\ldots,\lambda_\nu$ are the eigenvalues of J they will be resonant if one eigenvalue is an integer linear combination of two or more of the others. As terms of the form $\lambda_i - \sum$ (multiples of other eigenvalues) occurs in the denominator of the terms for the function h, the non-resonant condition is also known as the small divisor problem.

Conjugation results

The results on the existence of a solution to the conjugation equation depend on the eigenvalues of J and the degree of smoothness required from h. As J does not need to have any special symmetries, its eigenvalues will typically be complex numbers. When the eigenvalues of J are not in the unit circle, the dynamics near the fixed point x_0 of F is called *hyperbolic* and when the eigenvalues are on the unit circle and complex, the dynamics is called *elliptic*.

In the hyperbolic case the Hartman-Grobman theorem gives the conditions for the existence of a continuous function that maps the neighborhood of the fixed point of the map to the linear map $J \cdot x$. The hyperbolic case is also *structurally stable*. Small changes in the vector field will only produce small changes in the Poincaré map and these small changes will reflect in small changes in the position of the eigenvalues of J in the complex plane, implying that the map is still hyperbolic.

The Kolmogorov-Arnold-Moser (KAM) theorem gives the behavior near an elliptic point.

Bifurcation theory

When the evolution map Φ^t (or the vector field it is derived from) depends on a parameter μ, the structure of the phase space will also depend on this parameter. Small changes may produce no qualitative changes in the phase space until a special value μ_0 is reached. At this point the phase space changes qualitatively and the dynamical system is said to have gone through a bifurcation.

Bifurcation theory considers a structure in phase space (typically a fixed point, a periodic orbit, or an invariant torus) and studies its behavior as a function of the parameter μ. At the bifurcation point the structure may change its stability, split into new structures, or merge with other structures. By using Taylor series approximations of the maps and an

understanding of the differences that may be eliminated by a change of coordinates, it is possible to catalog the bifurcations of dynamical systems.

The bifurcations of a hyperbolic fixed point x_0 of a system family F_μ can be characterized by the eigenvalues of the first derivative of the system $DF_\mu(x_0)$ computed at the bifurcation point. For a map, the bifurcation will occur when there are eigenvalues of DF_μ on the unit circle. For a flow, it will occur when there are eigenvalues on the imaginary axis.

Some bifurcations can lead to very complicated structures in phase space. For example, the Ruelle–Takens scenario describes how a periodic orbit bifurcates into a torus and the torus into a strange attractor. In another example, Feigenbaum period-doubling describes how a stable periodic orbit goes through a series of period-doubling bifurcations.

Ergodic systems

In many dynamical systems it is possible to choose the coordinates of the system so that the volume (really a ν-dimensional volume) in phase space is invariant. This happens for mechanical systems derived from Newton's laws as long as the coordinates are the position and the momentum and the volume is measured in units of (position) × (momentum). The flow takes points of a subset A into the points $\Phi^t(A)$ and invariance of the phase space means that

$$\mathrm{vol}(A) = \mathrm{vol}(\Phi^t(A))\,.$$

In the Hamiltonian formalism, given a coordinate it is possible to derive the appropriate (generalized) momentum such that the associated volume is preserved by the flow. The volume is said to be computed by the Liouville measure.

In a Hamiltonian system not all possible configurations of position and momentum can be reached from an initial condition. Because of energy conservation, only the states with the same energy as the initial condition are accessible. The states with the same energy form an energy shell Ω, a sub-manifold of the phase space. The volume of the energy shell, computed using the Liouville measure, is preserved under evolution.

For systems where the volume is preserved by the flow, Poincaré discovered the recurrence theorem: Assume the phase space has a finite Liouville volume and let F be a phase space volume-preserving map and A a subset of the phase space. Then almost every point of A returns to A infinitely often. The Poincaré recurrence theorem was used by Zermelo to object to Boltzmann's derivation of the increase in entropy in a dynamical system of colliding atoms.

One of the questions raised by Boltzmann's work was the possible equality between time averages and space averages, what he called the ergodic hypothesis. The hypothesis states that the length of time a typical trajectory spends in a region A is vol(A)/vol(Ω).

The ergodic hypothesis turned out not to be the essential property needed for the development of statistical mechanics and a series of other ergodic-like properties were introduced to capture the relevant aspects of physical systems. Koopman approached the study of ergodic systems by the use of functional analysis. An observable a is a function that to each point of the phase space associates a number (say instantaneous pressure, or average height). The value of an observable can be computed at another time by using the evolution function φ^t. This introduces an operator U^t, the transfer operator,

$$(U^t a)(x) = a(\Phi^{-t}(x)) .$$

By studying the spectral properties of the linear operator U it becomes possible to classify the ergodic properties of Φ^t. In using the Koopman approach of considering the action of the flow on an observable function, the finite-dimensional nonlinear problem involving Φ^t gets mapped into an infinite-dimensional linear problem involving U.

The Liouville measure restricted to the energy surface Ω is the basis for the averages computed in equilibrium statistical mechanics. An average in time along a trajectory is equivalent to an average in space computed with the Boltzmann factor $\exp(-\beta H)$. This idea has been generalized by Sinai, Bowen, and Ruelle (SRB) to a larger class of dynamical systems that includes dissipative systems. SRB measures replace the Boltzmann factor and they are defined on attractors of chaotic systems.

Nonlinear dynamical systems and chaos

Simple nonlinear dynamical systems and even piecewise linear systems can exhibit a completely unpredictable behavior, which might seem to be random. (Remember that we are speaking of completely deterministic systems!). This seemingly unpredictable behavior has been called *chaos*. Hyperbolic systems are precisely defined dynamical systems that exhibit the properties ascribed to chaotic systems. In hyperbolic systems the tangent space perpendicular to a trajectory can be well separated into two parts: one with the points that converge towards the orbit (the *stable manifold*) and another of the points that diverge from the orbit (the *unstable manifold*).

This branch of mathematics deals with the long-term qualitative behavior of dynamical systems. Here, the focus is not on finding precise solutions to the equations defining the dynamical system (which is often hopeless), but rather to answer questions like "Will the system settle down to a steady state in the long term, and if so, what are the possible attractors?" or "Does the long-term behavior of the system depend on its initial condition?"

Note that the chaotic behavior of complicated systems is not the issue. Meteorology has been known for years to involve complicated—even chaotic—behavior. Chaos theory has been so surprising because chaos can be found within almost trivial systems. The logistic map is only a second-degree polynomial; the horseshoe map is piecewise linear.

Geometrical definition

A dynamical system is the tuple $\langle \mathcal{M}, f, \mathcal{T} \rangle$, with $\mathcal{M}$ a manifold (locally a Banach space or Euclidean space), $\mathcal{T}$ the domain for time (non-negative reals, the integers, ...) and *f* an evolution rule t→f^t (with $t \in \mathcal{T}$) such that f^t is a diffeomorphism of the manifold to itself. So, f is a mapping of the time-domain $\mathcal{T}$ into the space of diffeomorphisms of the manifold to itself. In other terms, f(t) is a diffeomorphism, for every time t in the domain $\mathcal{T}$.

Measure theoretical definition

A dynamical system may be defined formally, as a measure-preserving transformation of a sigma-algebra, the quadruplet (*X*,Σ,μ,τ). Here, *X* is a set, and Σ is a sigma-algebra on *X*, so that the pair (*X*,Σ) is a measurable space. μ is a finite measure on the sigma-algebra, so that the triplet (*X*,Σ,μ) is a probability space. A map $\tau : X \to X$ is said to be Σ-measurable if and only if, for every $\sigma \in \Sigma$, one has $\tau^{-1}\sigma \in \Sigma$. A map τ is said to **preserve the measure** if and only if, for every $\sigma \in \Sigma$, one has μ(τ^{-1}σ) = μ(σ). Combining the above, a map τ is said to be a **measure-preserving transformation of *X***, if it is a map from *X* to itself, it is Σ-measurable, and is measure-preserving. The quadruple (*X*,Σ,μ,τ), for such a τ, is then defined to be a **dynamical system**.

The map τ embodies the time evolution of the dynamical system. Thus, for discrete dynamical systems the iterates $\tau^n = \tau \circ \tau \circ \ldots \circ \tau$ for integer *n* are studied. For continuous dynamical systems, the map τ is understood to be a finite time evolution map and the construction is more complicated.

Examples of dynamical systems

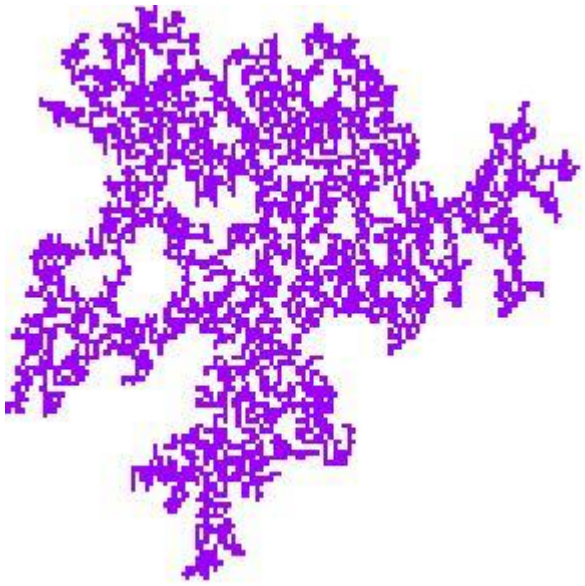
Internal links

- Arnold's cat map
- Baker's map is an example of a chaotic piecewise linear map
- Circle map
- Double pendulum
- Billiards and Outer Billiards
- Henon map
- Horseshoe map
- Irrational rotation
- List of chaotic maps
- Logistic map
- Lorenz system
- Rossler map

Chapter 5

Biological Engineering and Environmental Engineering

Biological engineering



Modeling of the spread of disease using Cellular Automata and Nearest Neighbor Interactions

Biological engineering, **biotechnological engineering** or **bioengineering** (including **biological systems engineering**) is the application of concepts and methods of physics, chemistry, and mathematics to solve problems in life sciences, using engineering's own analytical and synthetic methodologies. In this context, while traditional engineering applies physical and mathematical sciences to analyze, design and manufacture inanimate

tools, structures and processes, bioengineering uses the same sciences to study many aspects of living organisms. Usually it is used to analyze and solve problems related to human health.

Biological engineering is a science-based discipline founded upon the biological sciences in the same way that chemical engineering, electrical engineering, and mechanical engineering are based upon chemistry, electricity and magnetism and statics, respectively.

Biological engineering can be differentiated from its roots of pure biology or classical engineering in the following way. Biological studies often follow a reductionist approach in viewing a system on its smallest possible scale which naturally leads toward tools such as functional genomics. Engineering approaches, using classical design perspectives, are constructionist, building new devices, approaches, and technologies from component concepts. Biological engineering utilizes both of these methods in concert relying on reductionist approaches to define the fundamental units which are then commingled to generate something new. Although engineered biological systems have been used to manipulate information, construct materials, process chemicals, produce energy, provide food, and help maintain or enhance human health and our environment, our ability to quickly and reliably engineer biological systems that behave as expected remains less well developed than our mastery over mechanical and electrical systems.

The differentiation between biological engineering and overlap with Biomedical Engineering can be unclear, as many universities now use the terms "bioengineering" and "biomedical engineering" interchangeably . Some contend that Biological Engineering (like biotechnology) has a broader base which spans molecular methods (tends to emphasize the using of biological substances - applying engineering principles to molecular biology, biochemistry, microbiology, pharmacology, protein chemistry, cytology, immunology, neurobiology and neuroscience, cellular and tissue based methods (including devices and sensors), whole organisms (plants, animals), and up increasing length scales to ecosystems. Neither biological engineering nor biomedical engineering is wholly contained within the other, as there are non-biological products for medical needs and biological products for non-medical needs.

ABET , the U.S.-based accreditation board for engineering B.S. programs, makes a distinction between Biomedical Engineering and Biological Engineering; however, the differences are quite small. Biomedical engineers must have life science courses that include human physiology and have experience in performing measurements on living systems while biological engineers must have life science courses (which may or may not include physiology) and experience in making measurements not specifically on living systems. Foundational engineering courses are often the same and include thermodynamics, fluid and mechanical dynamics, kinetics, electronics, and materials properties.

The word bioengineering was coined by British scientist and broadcaster Heinz Wolff in 1954. The term bioengineering is also used to describe the use of vegetation in civil

engineering construction. The term bioengineering may also be applied to environmental modifications such as surface soil protection, slope stabilisation, watercourse and shoreline protection, windbreaks, vegetation barriers including noise barriers and visual screens, and the ecological enhancement of an area. The first biological engineering program was created at Mississippi State University in 1967, making it the first biological engineering curriculum in the United States. More recent programs have been launched at MIT and Utah State University .

Biological Engineers or *bioengineers* are engineers who use the principles of biology and the tools of engineering to create usable, tangible products. Biological Engineering employs knowledge and expertise from a number of pure and applied sciences, such as mass and heat transfer, kinetics, biocatalysts, biomechanics, bioinformatics, separation and purification processes, bioreactor design, surface science, fluid mechanics, thermodynamics, and polymer science. It is used in the design of medical devices, diagnostic equipment, biocompatible materials, renewable bioenergy, ecological engineering, and other areas that improve the living standards of societies.

In general, biological engineers attempt to either mimic biological systems in order to create products or modify and control biological systems so that they can replace, augment, or sustain chemical and mechanical processes. Bioengineers can apply their expertise to other applications of engineering and biotechnology, including genetic modification of plants and microorganisms, bioprocess engineering, and biocatalysis.

Because other engineering disciplines also address living organisms (e.g., prosthetics in mechanical engineering), the term biological engineering can be applied more broadly to include agricultural engineering and biotechnology. In fact, many old agricultural engineering departments in universities over the world have rebranded themselves as **agricultural and biological engineering** or **agricultural and biosystems engineering**. Biological engineering is also called bioengineering by some colleges and Biomedical engineering is called Bioengineering by others, and is a rapidly developing field with fluid categorization. The Main Fields of Bioengineering may be categorised as:

- **Bioprocess Engineering**: Bioprocess Design, Biocatalysis, Bioseparation, Bioinformatics, Bioenergy
- **Genetic Engineering**: Synthetic Biology, Horizontal gene transfer.
- **Cellular Engineering**: Cell Engineering, Tissue Culture Engineering, Metabolic engineering.
- **Biomedical Engineering**: Biomedical technology, Biomedical Diagnostics, Biomedical Therapy, Biomechanics, Biomaterials.

Environmental engineering

Environmental engineering is the application of science and engineering principles to improve the environment (air, water, and/or land resources), to provide healthy water, air, and land for human habitation and for other organisms, and to remediate polluted sites.

Environmental engineering involves waste water management and air pollution control, recycling, waste disposal, radiation protection, industrial hygiene, environmental sustainability, and public health issues as well as a knowledge of environmental engineering law. It also includes studies on the environmental impact of proposed construction projects.

Environmental engineers conduct hazardous-waste management studies to evaluate the significance of such hazards, advise on treatment and containment, and develop regulations to prevent mishaps. Environmental engineers also design municipal water supply and industrial wastewater treatment systems as well as address local and worldwide environmental issues such as the effects of acid rain, global warming, ozone depletion, water pollution and air pollution from automobile exhausts and industrial sources. At many universities, Environmental Engineering programs follow either the Department of Civil Engineering or The Department of Chemical Engineering at Engineering faculties. Environmental "civil" engineers focus on hydrology, water resources management, bioremediation, and water treatment plant design. Environmental "chemical" engineers, on the other hand, focus on environmental chemistry, advanced air and water treatment technologies and separation processes.

Additionally, engineers are more frequently obtaining specialized training in law (J.D.) and are utilizing their technical expertise in the practices of Environmental engineering law.. About four percent of environmental engineers go on to obtain Board Certification in their specialty area(s) of environmental engineering (Board Certified Environmental Engineer or BCEE)**.**

Most jurisdictions also impose licensing and registration requirements.

Development of environmental engineering

Ever since people first recognized that their health and well-being were related to the quality of their environment, they have applied thoughtful principles to attempt to improve the quality of their environment. The ancient Harappan civilization utilized early sewers in some cities. The Romans constructed aqueducts to prevent drought and to create a clean, healthful water supply for the metropolis of Rome. In the 15th century, Bavaria created laws restricting the development and degradation of alpine country that constituted the region's water supply

The field emerged as a separate environmental discipline during the middle third of the

20th century in response to widespread public concern about water and pollution and increasingly extensive environmental quality degradation. However, its roots extend back to early efforts in public health engineering. Modern environmental engineering began in London in the mid-19th century when Joseph Bazalgette designed the first major sewerage system that reduced the incidence of waterborne diseases such as cholera. The introduction of drinking water treatment and sewage treatment in industrialized countries reduced waterborne diseases from leading causes of death to rarities.

In many cases, as societies grew, actions that were intended to achieve benefits for those societies had longer-term impacts which reduced other environmental qualities. One example is the widespread application of DDT to control agricultural pests in the years following World War II. While the agricultural benefits were outstanding and crop yields increased dramatically, thus reducing world hunger substantially, and malaria was controlled better than it ever had been, numerous species were brought to the verge of extinction due to the impact of the DDT on their reproductive cycles. The story of DDT as vividly told in Rachel Carson's "Silent Spring" is considered to be the birth of the modern environmental movement and the development of the modern field of "environmental engineering."

Conservation movements and laws restricting public actions that would harm the environment have been developed by various societies for millennia. Notable examples are the laws decreeing the construction of sewers in London and Paris in the 19th century and the creation of the U.S. national park system in the early 20th century.

Scope of Environmental Engineering

Briefly speaking, the main task of environmental engineers is to protect public health by protecting (from further degradation), preserving (the present condition of), and enhancing the environment.

Environmental engineering is the application of science and engineering principles to the environment. Some consider environmental engineering to include the development of sustainable processes. There are several divisions of the field of environmental engineering.

Environmental impact assessment and mitigation

In this division, engineers and scientists use a systemic identification and evaluation process to assess the potential impacts of a proposed project , plans, programs, policies, or legislative actions upon the physical-chemical, biological, cultural, and socioeconomic components on environmental conditions. They apply scientific and engineering principles to evaluate if there are likely to be any adverse impacts to water quality, air quality, habitat quality, flora and fauna, agricultural capacity, traffic impacts, social impacts, ecological impacts, noise impacts, visual(landscape) impacts, etc. If impacts are expected, they then develop mitigation measures to limit or prevent such impacts. An example of a mitigation measure would be the creation of wetlands in a nearby location

to mitigate the filling in of wetlands necessary for a road development if it is not possible to reroute the road.

The practice of environmental assessment was intitiated on January 1, 1970, the effective date of the National Environmental Policy Act (NEPA) in the United States. Since that time, more than 100 developing and developed nations either have planned specific analogous laws or have adopted procedure used elsewhere. NEPA is applicable to all federal agencies in the United States.

Water supply and treatment

Engineers and scientists work to secure water supplies for potable and agricultural use. They evaluate the water balance within a watershed and determine the available water supply, the water needed for various needs in that watershed, the seasonal cycles of water movement through the watershed and they develop systems to store, treat, and convey water for various uses. Water is treated to achieve water quality objectives for the end uses. In the case of potable water supply, water is treated to minimize the risk of infectious disease transmission, the risk of non-infectious illness, and to create a palatable water flavor. Water distribution systems are designed and built to provide adequate water pressure and flow rates to meet various end-user needs such as domestic use, fire suppression, and irrigation.

Wastewater conveyance and treatment

Water pollution

Most urban and many rural areas no longer discharge human waste directly to the land through outhouse, septic, and/or honey bucket systems, but rather deposit such waste into water and convey it from households via sewer systems. Engineers and scientists develop collection and treatment systems to carry this waste material away from where people live and produce the waste and discharge it into the environment. In developed countries, substantial resources are applied to the treatment and detoxification of this waste before it is discharged into a river, lake, or ocean system. Developing nations are striving to obtain the resources to develop such systems so that they can improve water quality in their surface waters and reduce the risk of water-borne infectious disease.

Sewage treatment plant, Australia.

There are numerous wastewater treatment technologies. A wastewater treatment train can consist of a primary clarifier system to remove solid and floating materials, a secondary treatment system consisting of an aeration basin followed by flocculation and sedimentation or an activated sludge system and a secondary clarifier, a tertiary biological nitrogen removal system, and a final disinfection process. The aeration basin/activated sludge system removes organic material by growing bacteria (activated sludge). The secondary clarifier removes the activated sludge from the water. The tertiary system, although not always included due to costs, is becoming more prevalent to remove nitrogen and phosphorus and to disinfect the water before discharge to a surface water stream or ocean outfall.

Air quality management

Engineers apply scientific and engineering principles to the design of manufacturing and combustion processes to reduce air pollutant emissions to acceptable levels. Scrubbers, electrostatic precipitators, catalytic converters, and various other processes are utilized to remove particulate matter, nitrogen oxides, sulfur oxides, volatile organic compounds (VOC), reactive organic gases (ROG) and other air pollutants from flue gases and other sources prior to allowing their emission to the atmosphere.

Scientists have developed air pollution dispersion models to evaluate the concentration of a pollutant at a receptor or the impact on overall air quality from vehicle exhausts and industrial flue gas stack emissions.

To some extent, this field overlaps the desire to decrease carbon dioxide and other greenhouse gas emissions from combustion processes**.**

Other applications

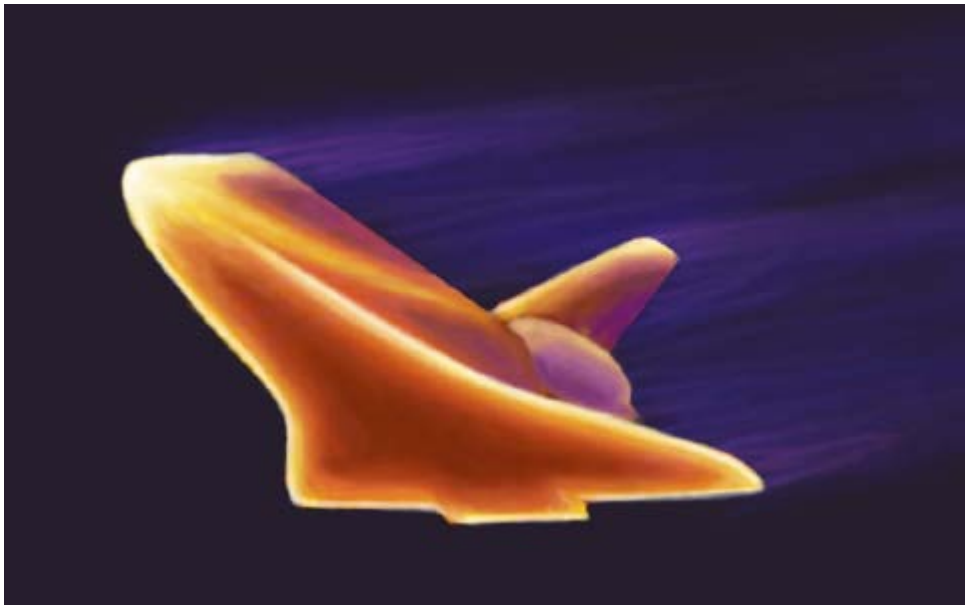
- Environmental policy and regulation development
- Contaminated land management and site remediation
- Environment, Health and Safety
- Hazardous waste management
- Natural resource management
- Noise pollution
- Risk assessment
- Solid waste management

Prominent environmental engineers

- Robert A. Gearheart
- Paul V. Roberts
- Abel Wolman

Chapter 6

Materials Science



Simulation of the outside of the Space Shuttle as it heats up to over 1,500 °C (2,730 °F) during re-entry into the Earth's atmosphere

Materials science is an interdisciplinary field applying the properties of matter to various areas of science and engineering. This scientific field investigates the relationship between the structure of materials at atomic or molecular scales and their macroscopic properties. It incorporates elements of applied physics and chemistry. With significant media attention focused on nanoscience and nanotechnology in recent years, materials science has been propelled to the forefront at many universities. It is also an important part of forensic engineering and failure analysis. Materials science also deals with *fundamental properties* and *characteristics* of materials.

History

The material of choice of a given era is often its defining point. Phrases such as Stone Age, Bronze Age, and Steel Age are good examples. Originally deriving from the manufacture of ceramics and its putative derivative metallurgy, materials science is one of the oldest forms of engineering and applied science. Modern materials science evolved directly from metallurgy, which itself evolved from mining and (likely) ceramics and the use of fire. A major breakthrough in the understanding of materials occurred in the late 19th century, when the American scientist Josiah Willard Gibbs demonstrated that the thermodynamic properties related to atomic structure in various phases are related to the physical properties of a material. Important elements of modern materials science are a product of the space race: the understanding and engineering of the metallic alloys, and silica and carbon materials, used in the construction of space vehicles enabling the exploration of space. Materials science has driven, and been driven by, the development of revolutionary technologies such as plastics, semiconductors, and biomaterials.

Before the 1960s (and in some cases decades after), many *materials science* departments were named *metallurgy* departments, from a 19th and early 20th century emphasis on metals. The field has since broadened to include every class of materials, including: ceramics, polymers, semiconductors, magnetic materials, medical implant materials and biological materials (materiomics).

Fundamentals

The basis of materials science involves relating the desired properties and relative performance of a material in a certain application to the structure of the atoms and phases in that material through characterization. The major determinants of the structure of a material and thus of its properties are its constituent chemical elements and the way in which it has been processed into its final form. These characteristics, taken together and related through the laws of thermodynamics, govern a material's microstructure, and thus its properties.

The manufacture of a perfect crystal of a material is currently physically impossible. Instead materials scientists manipulate the defects in crystalline materials such as precipitates, grain boundaries (Hall-Petch relationship), interstitial atoms, vacancies or substitutional atoms, to create materials with the desired properties.

Not all materials have a regular crystal structure. Polymers display varying degrees of crystallinity, and many are completely non-crystalline. Glasses, some ceramics, and many natural materials are amorphous, not possessing any long-range order in their atomic arrangements. The study of polymers combines elements of chemical and statistical thermodynamics to give thermodynamic, as well as mechanical, descriptions of physical properties.

In addition to industrial interest, materials science has gradually developed into a field which provides tests for condensed matter or solid state theories. New physics emerge because of the diverse new material properties which need to be explained.

Materials in industry

Radical materials advances can drive the creation of new products or even new industries, but stable industries also employ materials scientists to make incremental improvements and troubleshoot issues with currently used materials. Industrial applications of materials science include materials design, cost-benefit tradeoffs in industrial production of materials, processing techniques (casting, rolling, welding, ion implantation, crystal growth, thin-film deposition, sintering, glassblowing, etc.), and analytical techniques (characterization techniques such as electron microscopy, x-ray diffraction, calorimetry, nuclear microscopy (HEFIB), Rutherford backscattering, neutron diffraction, small-angle X-ray scattering (SAXS), etc.

Besides material characterization, the material scientist/engineer also deals with the extraction of materials and their conversion into useful forms. Thus ingot casting, foundry techniques, blast furnace extraction, and electrolytic extraction are all part of the required knowledge of a metallurgist/engineer. Often the presence, absence or variation of minute quantities of secondary elements and compounds in a bulk material will have a great impact on the final properties of the materials produced, for instance, steels are classified based on 1/10 and 1/100 weight percentages of the carbon and other alloying elements they contain. Thus, the extraction and purification techniques employed in the extraction of iron in the blast furnace will have an impact of the quality of steel that may be produced.

The overlap between physics and materials science has led to the offshoot field of *materials physics*, which is concerned with the physical properties of materials. The approach is generally more macroscopic and applied than in condensed matter physics.

Metal alloys

The study of metal alloys is a significant part of materials science. Of all the metallic alloys in use today, the alloys of iron (steel, stainless steel, cast iron, tool steel, alloy steels) make up the largest proportion both by quantity and commercial value. Iron alloyed with various proportions of carbon gives low, mid and high carbon steels. For the steels, the hardness and tensile strength of the steel is directly related to the amount of carbon present, with increasing carbon levels also leading to lower ductility and toughness. The addition of silicon and graphitization will produce cast irons (although some cast irons are made precisely with no graphitization). The addition of chromium, nickel and molybdenum to carbon steels (more than 10%) gives us stainless steels.

Other significant metallic alloys are those of aluminium, titanium, copper and magnesium. Copper alloys have been known for a long time (since the Bronze Age), while the alloys of the other three metals have been relatively recently developed. Due to

the chemical reactivity of these metals, the electrolytic extraction processes required were only developed relatively recently. The alloys of aluminium, titanium and magnesium are also known and valued for their high strength-to-weight ratios and, in the case of magnesium, their ability to provide electromagnetic shielding. These materials are ideal for situations where high strength-to-weight ratios are more important than bulk cost, such as in the aerospace industry and certain automotive engineering applications.

Polymers

Polymers are also an important part of materials science. Polymers are the raw materials (the resins) used to make what we commonly call plastics. Plastics are really the final product, created after one or more polymers or additives have been added to a resin during processing, which is then shaped into a final form. Polymers which have been around, and which are in current widespread use, include polyethylene, polypropylene, PVC, polystyrene, nylons, polyesters, acrylics, polyurethanes, and polycarbonates. Plastics are generally classified as "commodity", "specialty" and "engineering" plastics.

PVC (polyvinyl-chloride) is widely used, inexpensive, and annual production quantities are large. It lends itself to an incredible array of applications, from artificial leather to electrical insulation and cabling, packaging and containers. Its fabrication and processing are simple and well-established. The versatility of PVC is due to the wide range of plasticisers and other additives that it accepts. The term "additives" in polymer science refers to the chemicals and compounds added to the polymer base to modify its material properties.

Polycarbonate would be normally considered an engineering plastic (other examples include PEEK, ABS). Engineering plastics are valued for their superior strengths and other special material properties. They are usually not used for disposable applications, unlike commodity plastics.

Specialty plastics are materials with unique characteristics, such as ultra-high strength, electrical conductivity, electro-fluorescence, high thermal stability, etc.

The dividing lines between the various types of plastics is not based on material but rather on their properties and applications. For instance, polyethylene (PE) is a cheap, low friction polymer commonly used to make disposable shopping bags and trash bags, and is considered a commodity plastic, whereas Medium-Density Polyethylene MDPE is used for underground gas and water pipes, and another variety called Ultra-high Molecular Weight Polyethylene UHMWPE is an engineering plastic which is used extensively as the glide rails for industrial equipment and the low-friction socket in implanted hip joints.

Ceramics and glasses

Composite materials

Another application of material science in industry is the making of composite materials. Composite materials are structured materials composed of two or more macroscopic phases. Applications range from structural elements such as steel-reinforced concrete, to the thermally insulative tiles which play a key and integral role in NASA's Space Shuttle thermal protection system which is used to protect the surface of the shuttle from the heat of re-entry into the Earth's atmosphere. One example is Reinforced Carbon-Carbon (RCC), The light gray material which withstands re-entry temperatures up to 1510 °C (2750 °F) and protects the Space Shuttle's wing leading edges and nose cap. RCC is a laminated composite material made from graphite rayon cloth and impregnated with a phenolic resin. After curing at high temperature in an autoclave, the laminate is pyrolized to convert the resin to carbon, impregnated with furfural alcohol in a vacuum chamber, and cured/pyrolized to convert the furfural alcohol to carbon. In order to provide oxidation resistance for reuse capability, the outer layers of the RCC are converted to silicon carbide.

Other examples can be seen in the "plastic" casings of television sets, cell-phones and so on. These plastic casings are usually a composite material made up of a thermoplastic matrix such as acrylonitrile-butadiene-styrene (ABS) in which calcium carbonate chalk, talc, glass fibres or carbon fibres have been added for added strength, bulk, or electro-static dispersion. These additions may be referred to as reinforcing fibres, or dispersants, depending on their purpose.

Classes of materials

Materials science encompasses various classes of materials, each of which may constitute a separate field. Materials are sometimes classified by the type of bonding present between the atoms:

1. Ionic crystals
2. Covalent crystals
3. Metals
4. Intermetallics
5. Semiconductors
6. Polymers
7. Composite materials
8. Vitreous materials

Overview

- Nanotechnology – rigorously, the study of materials where the effects of quantum confinement, the Gibbs-Thomson effect, or any other effect only present at the nanoscale is the defining property of the material; but more commonly, it is the

creation and study of materials whose defining structural properties are anywhere from less than a nanometer to one hundred nanometers in scale, such as molecularly engineered materials.

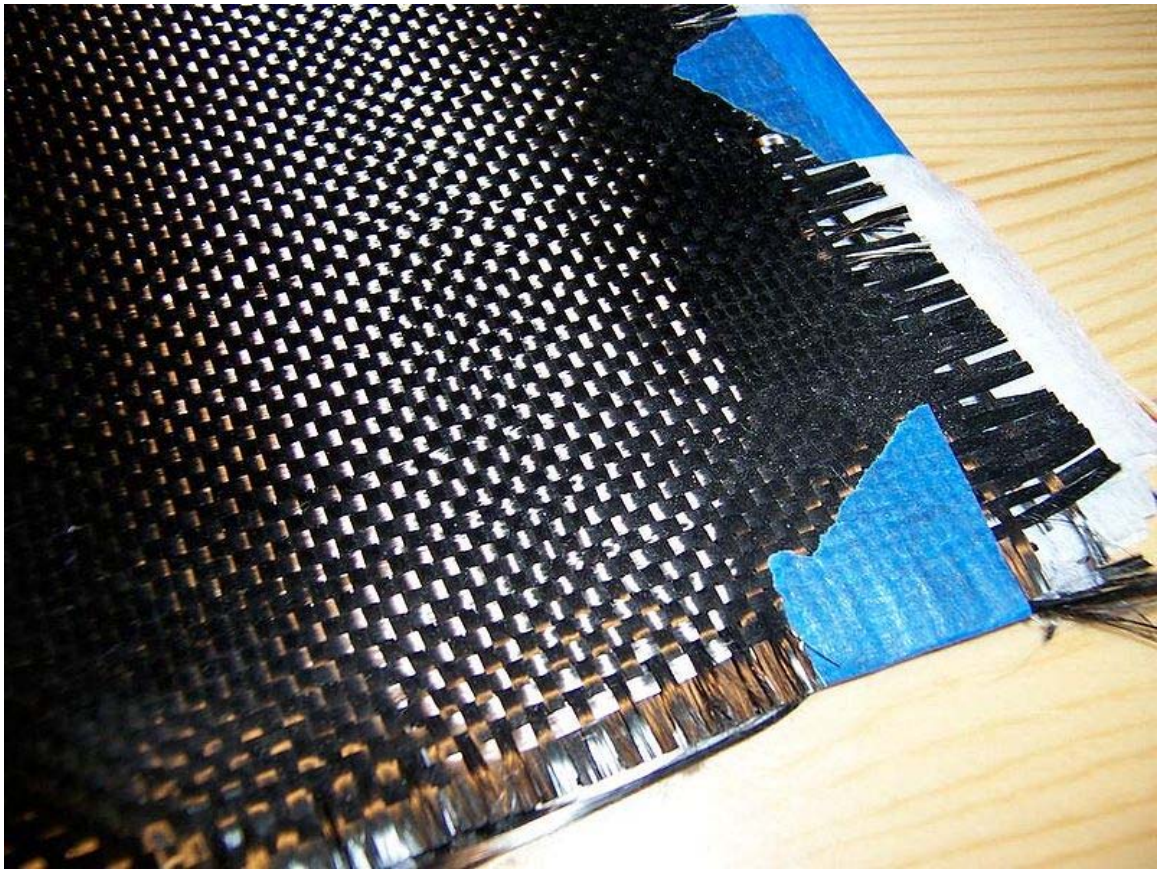
- Microtechnology - study of materials and processes and their interaction, allowing microfabrication of structures of micrometric dimensions, such as MicroElectroMechanical Systems (MEMS).
- Crystallography – the study of how atoms in a solid fill space, the defects associated with crystal structures such as grain boundaries and dislocations, and the characterization of these structures and their relation to physical properties.
- Materials Characterization – such as diffraction with x-rays, electrons, or neutrons, and various forms of spectroscopy and chemical analysis such as Raman spectroscopy, energy-dispersive spectroscopy (EDS), chromatography, thermal analysis, electron microscope analysis, etc., in order to understand and define the properties of materials.

Si_3N_4 ceramic bearing parts

- Metallurgy – the study of metals and their alloys, including their extraction, microstructure and processing.
- Biomaterials – materials that are derived from and/or used with biological systems.

- Electronic and magnetic materials – materials such as semiconductors used to create integrated circuits, storage media, sensors, and other devices.
- Tribology – the study of the wear of materials due to friction and other factors.
- Surface science/Catalysis – interactions and structures between solid-gas solid-liquid or solid-solid interfaces.
- Ceramography – the study of the microstructures of high-temperature materials and refractories, including structural ceramics such as RCC, polycrystalline silicon carbide and transformation toughened ceramics

Some practitioners often consider rheology a sub-field of materials science, because it can cover any material that flows. However, modern rheology typically deals with non-Newtonian fluid dynamics, so it is often considered a sub-field of continuum mechanics.



A cloth of woven carbon fiber filaments is commonly used for reinforcement in composite materials.

- Glass Science – any non-crystalline material including inorganic glasses, vitreous metals and non-oxide glasses.
- Forensic engineering – the study of how products fail, and the vital role of the materials of construction
- Forensic materials engineering – the study of material failure, and the light it sheds on how engineers specify materials in their product

- Textile Reinforced Materials - materials in the form of ceramic or concrete are reinforced with a primarily woven or non-woven textile structure to impose high strength with comparatively more flexibility to withstand vibrations and sudden jerks.

Primary topics

- Thermodynamics, statistical mechanics, and physical chemistry, for phase equilibrium conditions, phase diagrams of materials systems (multi-phase, multi-component, reacting and non-reacting systems)
- Phase transformation kinetics, for the kinetics of phase transformations (with particular emphasis on solid-solid phase transitions)
- Transport phenomena for the transport of heat, mass, and momentum in materials processing.
- Crystallography, quantum chemistry or quantum physics, for the structure (symmetry and defects) and bonding in materials (e.g., ionic, metallic, covalent, and van der Waals bonding)
- Mechanical behavior of materials, to understand the mechanical properties of materials, defects and their propagation, and their behavior under static, dynamic, and cyclic loads
- Electronic properties of materials, and solid-state physics, for the understanding of the electronic, thermal, magnetic, and optical properties of materials
- Diffraction and wave mechanics, for the science behind characterization systems, e.g., transmission electron microscopy (TEM)

Household items made of various kinds of plastic.

- Polymer properties, synthesis, and characterization, for a specialized understanding of how polymers behave, how they are made, and how they are characterized; exciting applications of polymers include liquid crystal displays (LCDs, the displays found in most cell-phones, cameras, and iPods), novel photovoltaic devices based on semiconductor polymers (which, unlike the traditional silicon solar panels, are flexible and cheap to manufacture, albeit with lower efficiency), and membranes for room-temperature fuel cells (as proton exchange membranes) and filtration systems in the environmental and biomedical fields
- Biomaterials, physiology, biomechanics, biochemistry, for a specialized understanding of how materials integrate into biological systems, e.g., through materiomics
- Semiconductor materials and semiconductor devices, for a specialized understanding of the advanced processes used in industry (e.g. crystal growth techniques, thin-film deposition, ion implantation, photolithography), their properties, and their integration in electronic devices
- Alloying, corrosion, and thermal or mechanical processing, for a specialized treatment of metallurgical materials—with applications ranging from aerospace and industrial equipment to the civil industries

Professional organizations

- Materials Research Society, MRS
- European Materials Research Society, EMRS
- ASM International
- The Minerals, Metals, & Materials Society, TMS
- Materials Australia
- American Ceramic Society, ACerS
- NACE International
- The American Institute of Mining, Metallurgical, and Petroleum Engineers, AIME
- Society for the Advancement of Material and Process Engineering, SAMPE
- The Institute of Materials, Minerals and Mining, IOM^3
- Alpha Sigma Mu, AΣM
- Central European Institute of Technology, CEITEC

Chapter 7

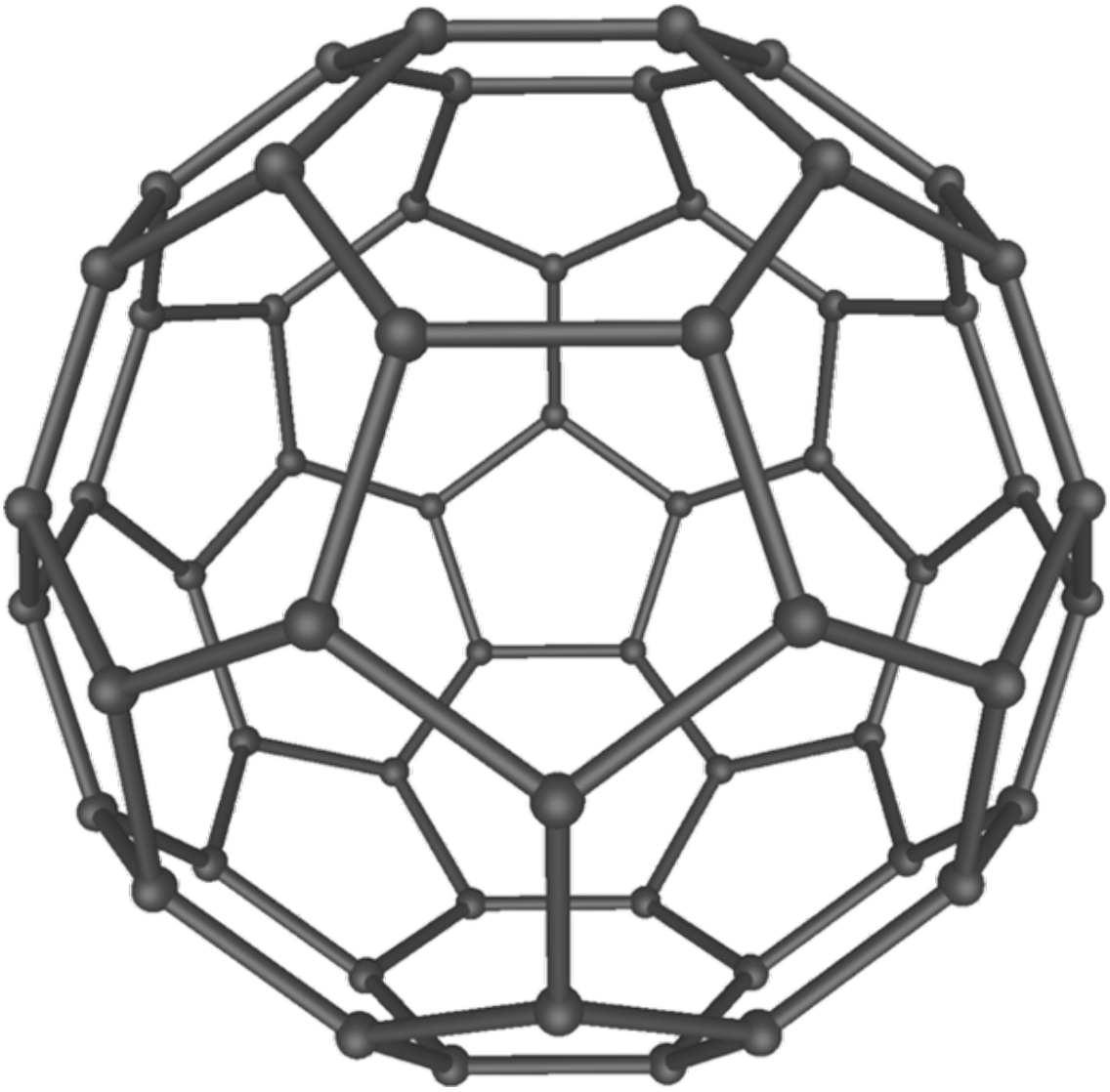
Nanotechnology

Nanotechnology (sometimes shortened to "**nanotech**") is the study of manipulating matter on an atomic and molecular scale. Generally, nanotechnology deals with structures sized between 1 to 100 nanometre in at least one dimension, and involves developing materials or devices possessing at least one dimension within that size. Quantum mechanical effects are very important at this scale, which is in the quantum realm.

Nanotechnology is very diverse, ranging from extensions of conventional device physics to completely new approaches based upon molecular self-assembly, from developing new materials with dimensions on the nanoscale to investigating whether we can directly control matter on the atomic scale.

There is much debate on the future implications of nanotechnology. Nanotechnology may be able to create many new materials and devices with a vast range of applications, such as in medicine, electronics, biomaterials and energy production. On the other hand, nanotechnology raises many of the same issues as any new technology, including concerns about the toxicity and environmental impact of nanomaterials, and their potential effects on global economics, as well as speculation about various doomsday scenarios. These concerns have led to a debate among advocacy groups and governments on whether special regulation of nanotechnology is warranted.

Origins



Buckminsterfullerene C_{60}, also known as the buckyball, is a representative member of the carbon structures known as fullerenes. Members of the fullerene family are a major subject of research falling under the nanotechnology umbrella.

The first use of the concepts found in 'nano-technology' (but pre-dating use of that name) was in "There's Plenty of Room at the Bottom", a talk given by physicist Richard Feynman at an American Physical Society meeting at California Institute of Technology (Caltech) on December 29, 1959. Feynman described a process by which the ability to manipulate individual atoms and molecules might be developed, using one set of precise tools to build and operate another proportionally smaller set, and so on down to the needed scale. In the course of this, he noted, scaling issues would arise from the changing magnitude of various physical phenomena: gravity would become less important, surface tension and van der Waals attraction would become increasingly more significant, etc. This basic idea appeared plausible, and exponential assembly enhances it with parallelism

to produce a useful quantity of end products. The term "nanotechnology" was defined by Tokyo University of Science Professor Norio Taniguchi in a 1974 paper as follows: "'Nano-technology' mainly consists of the processing of, separation, consolidation, and deformation of materials by one atom or by one molecule." In the 1980s the basic idea of this definition was explored in much more depth by Dr. K. Eric Drexler, who promoted the technological significance of nano-scale phenomena and devices through speeches and the books *Engines of Creation: The Coming Era of Nanotechnology* (1986) and *Nanosystems: Molecular Machinery, Manufacturing, and Computation*, and so the term acquired its current sense. *Engines of Creation* is considered the first book on the topic of nanotechnology. Nanotechnology and nanoscience got started in the early 1980s with two major developments; the birth of cluster science and the invention of the scanning tunneling microscope (STM). This development led to the discovery of fullerenes in 1985 and carbon nanotubes a few years later. In another development, the synthesis and properties of semiconductor nanocrystals was studied; this led to a fast increasing number of metal and metal oxide nanoparticles and quantum dots. The atomic force microscope (AFM or SFM) was invented six years after the STM was invented. In 2000, the United States National Nanotechnology Initiative was founded to coordinate Federal nanotechnology research and development and is evaluated by the President's Council of Advisors on Science and Technology.

Fundamental concepts

Nanotechnology is the engineering of functional systems at the molecular scale. This covers both current work and concepts that are more advanced. In its original sense, nanotechnology refers to the projected ability to construct items from the bottom up, using techniques and tools being developed today to make complete, high performance products.

One nanometer (nm) is one billionth, or 10^{-9}, of a meter. By comparison, typical carbon-carbon bond lengths, or the spacing between these atoms in a molecule, are in the range 0.12–0.15 nm, and a DNA double-helix has a diameter around 2 nm. On the other hand, the smallest cellular life-forms, the bacteria of the genus Mycoplasma, are around 200 nm in length. By convention, nanotechnology is taken as the scale range 1 to 100 nm following the definition used by the National Nanotechnology Initiative in the US. The lower limit is set by the size of atoms (hydrogen has the smallest atoms, which are approximately a quarter of a nm diameter) since nanotechnology must build its devices from atoms and molecules. The upper limit is more or less arbitrary but is around the size that phenomena not observed in larger structures start to become apparent and can be made use of in the nano device. These new phenomena make nanotechnology distinct from devices which are merely miniaturised versions of an equivalent macroscopic device; such devices are on a larger scale and come under the description of microtechnology.

To put that scale in another context, the comparative size of a nanometer to a meter is the same as that of a marble to the size of the earth. Or another way of putting it: a nanometer

is the amount an average man's beard grows in the time it takes him to raise the razor to his face.

Two main approaches are used in nanotechnology. In the "bottom-up" approach, materials and devices are built from molecular components which assemble themselves chemically by principles of molecular recognition. In the "top-down" approach, nano-objects are constructed from larger entities without atomic-level control.

Areas of physics such as nanoelectronics, nanomechanics, nanophotonics and nanoionics have evolved during the last few decades to provide a basic scientific foundation of nanotechnology.

Larger to smaller: a materials perspective

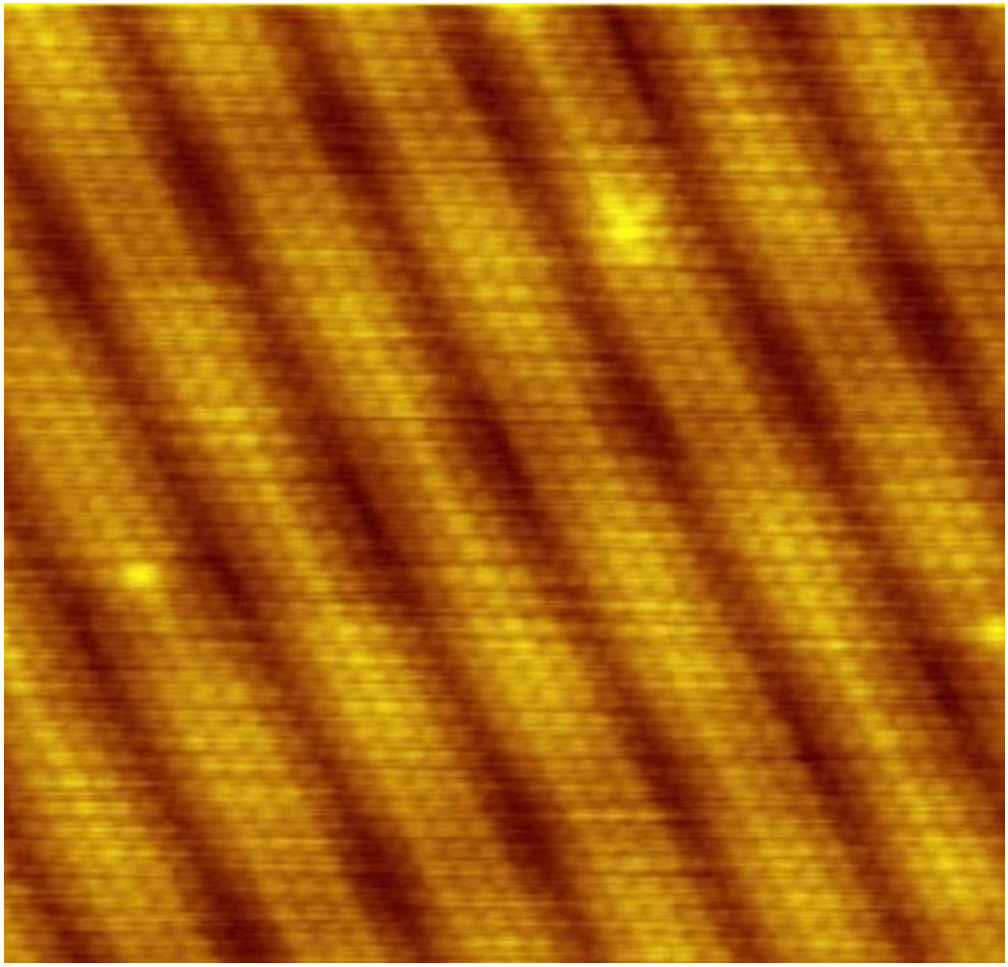


Image of reconstruction on a clean Gold(100) surface, as visualized using scanning tunneling microscopy. The positions of the individual atoms composing the surface are visible.

A number of physical phenomena become pronounced as the size of the system decreases. These include statistical mechanical effects, as well as quantum mechanical

effects, for example the “quantum size effect” where the electronic properties of solids are altered with great reductions in particle size. This effect does not come into play by going from macro to micro dimensions. However, quantum effects become dominant when the nanometer size range is reached, typically at distances of 100 nanometers or less, the so called quantum realm. Additionally, a number of physical (mechanical, electrical, optical, etc.) properties change when compared to macroscopic systems. One example is the increase in surface area to volume ratio altering mechanical, thermal and catalytic properties of materials. Diffusion and reactions at nanoscale, nanostructures materials and nanodevices with fast ion transport are generally referred to nanoionics. *Mechanical* properties of nanosystems are of interest in the nanomechanics research. The catalytic activity of nanomaterials also opens potential risks in their interaction with biomaterials.

Materials reduced to the nanoscale can show different properties compared to what they exhibit on a macroscale, enabling unique applications. For instance, opaque substances become transparent (copper); stable materials turn combustible (aluminum); insoluble materials become soluble (gold). A material such as gold, which is chemically inert at normal scales, can serve as a potent chemical catalyst at nanoscales. Much of the fascination with nanotechnology stems from these quantum and surface phenomena that matter exhibits at the nanoscale.

Simple to complex: a molecular perspective

Modern synthetic chemistry has reached the point where it is possible to prepare small molecules to almost any structure. These methods are used today to manufacture a wide variety of useful chemicals such as pharmaceuticals or commercial polymers. This ability raises the question of extending this kind of control to the next-larger level, seeking methods to assemble these single molecules into supramolecular assemblies consisting of many molecules arranged in a well defined manner.

These approaches utilize the concepts of molecular self-assembly and/or supramolecular chemistry to automatically arrange themselves into some useful conformation through a bottom-up approach. The concept of molecular recognition is especially important: molecules can be designed so that a specific configuration or arrangement is favored due to non-covalent intermolecular forces. The Watson–Crick basepairing rules are a direct result of this, as is the specificity of an enzyme being targeted to a single substrate, or the specific folding of the protein itself. Thus, two or more components can be designed to be complementary and mutually attractive so that they make a more complex and useful whole.

Such bottom-up approaches should be capable of producing devices in parallel and be much cheaper than top-down methods, but could potentially be overwhelmed as the size and complexity of the desired assembly increases. Most useful structures require complex and thermodynamically unlikely arrangements of atoms. Nevertheless, there are many examples of self-assembly based on molecular recognition in biology, most notably Watson–Crick basepairing and enzyme-substrate interactions. The challenge for

nanotechnology is whether these principles can be used to engineer new constructs in addition to natural ones.

Molecular nanotechnology: a long-term view

Molecular nanotechnology, sometimes called molecular manufacturing, describes engineered nanosystems (nanoscale machines) operating on the molecular scale. Molecular nanotechnology is especially associated with the molecular assembler, a machine that can produce a desired structure or device atom-by-atom using the principles of mechanosynthesis. Manufacturing in the context of productive nanosystems is not related to, and should be clearly distinguished from, the conventional technologies used to manufacture nanomaterials such as carbon nanotubes and nanoparticles.

When the term "nanotechnology" was independently coined and popularized by Eric Drexler (who at the time was unaware of an earlier usage by Norio Taniguchi) it referred to a future manufacturing technology based on molecular machine systems. The premise was that molecular scale biological analogies of traditional machine components demonstrated molecular machines were possible: by the countless examples found in biology, it is known that sophisticated, stochastically optimised biological machines can be produced.

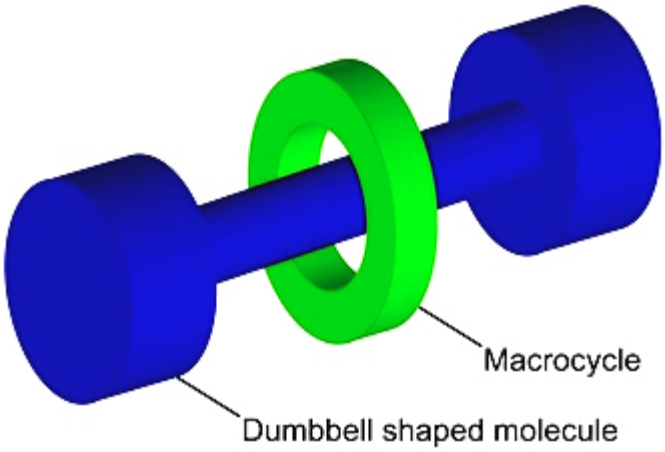
It is hoped that developments in nanotechnology will make possible their construction by some other means, perhaps using biomimetic principles. However, Drexler and other researchers have proposed that advanced nanotechnology, although perhaps initially implemented by biomimetic means, ultimately could be based on mechanical engineering principles, namely, a manufacturing technology based on the mechanical functionality of these components (such as gears, bearings, motors, and structural members) that would enable programmable, positional assembly to atomic specification. The physics and engineering performance of exemplar designs were analyzed in Drexler's book *Nanosystems*.

In general it is very difficult to assemble devices on the atomic scale, as all one has to position atoms on other atoms of comparable size and stickiness. Another view, put forth by Carlo Montemagno, is that future nanosystems will be hybrids of silicon technology and biological molecular machines. Yet another view, put forward by the late Richard Smalley, is that mechanosynthesis is impossible due to the difficulties in mechanically manipulating individual molecules.

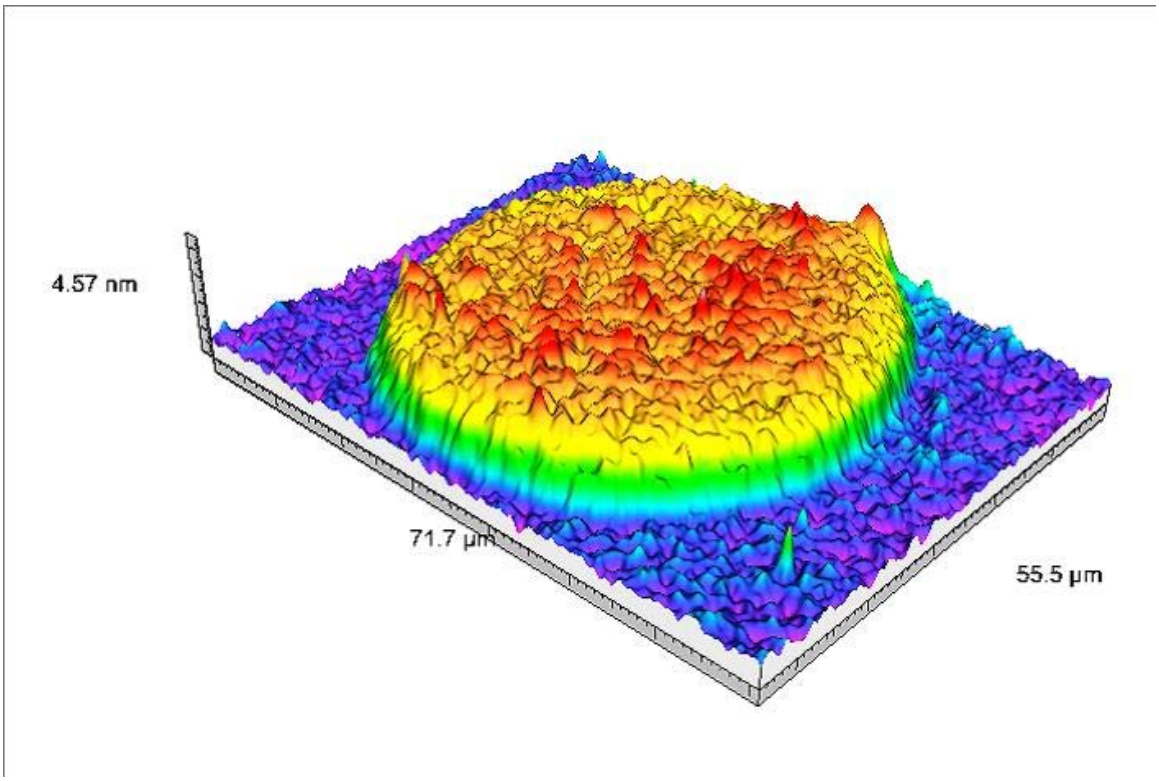
This led to an exchange of letters in the ACS publication Chemical & Engineering News in 2003. Though biology clearly demonstrates that molecular machine systems are possible, non-biological molecular machines are today only in their infancy. Leaders in research on non-biological molecular machines are Dr. Alex Zettl and his colleagues at Lawrence Berkeley Laboratories and UC Berkeley. They have constructed at least three distinct molecular devices whose motion is controlled from the desktop with changing voltage: a nanotube nanomotor, a molecular actuator, and a nanoelectromechanical relaxation oscillator.

An experiment indicating that positional molecular assembly is possible was performed by Ho and Lee at Cornell University in 1999. They used a scanning tunneling microscope to move an individual carbon monoxide molecule (CO) to an individual iron atom (Fe) sitting on a flat silver crystal, and chemically bound the CO to the Fe by applying a voltage.

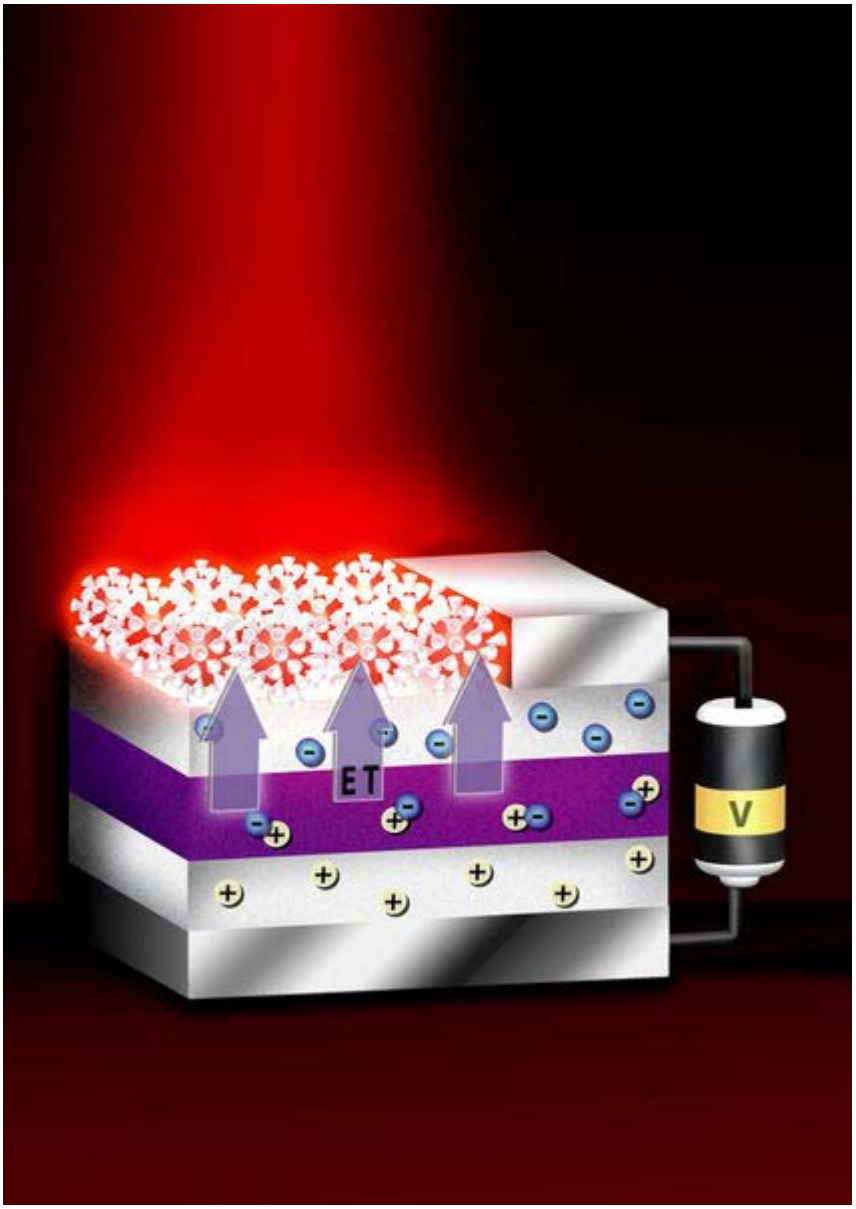
Current research



Graphical representation of a rotaxane, useful as a molecular switch.



Sarfus image of a DNA biochip elaborated by bottom-up approach.



This device transfers energy from nano-thin layers of quantum wells to nanocrystals above them, causing the nanocrystals to emit visible light.

Nanomaterials

The nanomaterials field includes subfields which develop or study materials having unique properties arising from their nanoscale dimensions.

- Interface and colloid science has given rise to many materials which may be useful in nanotechnology, such as carbon nanotubes and other fullerenes, and various nanoparticles and nanorods. Nanomaterials with fast ion transport are related also to nanoionics and nanoelectronics.

- Nanoscale materials can also be used for bulk applications; most present commercial applications of nanotechnology are of this flavor.
- Progress has been made in using these materials for medical applications.
- Nanoscale materials are sometimes used in solar cells which combats the cost of traditional Silicon solar cells
- Development of applications incorporating semiconductor nanoparticles to be used in the next generation of products, such as display technology, lighting, solar cells and biological imaging.

Bottom-up approaches

These seek to arrange smaller components into more complex assemblies.

- DNA nanotechnology utilizes the specificity of Watson–Crick basepairing to construct well-defined structures out of DNA and other nucleic acids.
- Approaches from the field of "classical" chemical synthesis also aim at designing molecules with well-defined shape (e.g. bis-peptides).
- More generally, molecular self-assembly seeks to use concepts of supramolecular chemistry, and molecular recognition in particular, to cause single-molecule components to automatically arrange themselves into some useful conformation.
- Atomic force microscope tips can be used as a nanoscale "write head" to deposit a chemical upon a surface in a desired pattern in a process called dip pen nanolithography. This technique fits into the larger subfield of nanolithography.

Top-down approaches

These seek to create smaller devices by using larger ones to direct their assembly.

- Many technologies that descended from conventional solid-state silicon methods for fabricating microprocessors are now capable of creating features smaller than 100 nm, falling under the definition of nanotechnology. Giant magnetoresistance-based hard drives already on the market fit this description, as do atomic layer deposition (ALD) techniques. Peter Grünberg and Albert Fert received the Nobel Prize in Physics in 2007 for their discovery of Giant magnetoresistance and contributions to the field of spintronics.
- Solid-state techniques can also be used to create devices known as nanoelectromechanical systems or NEMS, which are related to microelectromechanical systems or MEMS.
- Focused ion beams can directly remove material, or even deposit material when suitable pre-cursor gasses are applied at the same time. For example, this technique is used routinely to create sub-100 nm sections of material for analysis in Transmission electron microscopy.
- Atomic force microscope tips can be used as a nanoscale "write head" to deposit a resist, which is then followed by a etching process to remove material in a top-down method.

Functional approaches

These seek to develop components of a desired functionality without regard to how they might be assembled.

- Molecular electronics seeks to develop molecules with useful electronic properties. These could then be used as single-molecule components in a nanoelectronic device.
- Synthetic chemical methods can also be used to create synthetic molecular motors, such as in a so-called nanocar.

Biomimetic approaches

- Bionics or biomimicry seeks to apply biological methods and systems found in nature, to the study and design of engineering systems and modern technology. Biomineralization is one example of the systems studied.

- Bionanotechnology the use of biomolecules for applications in nanotechnology, including use of viruses.

Speculative

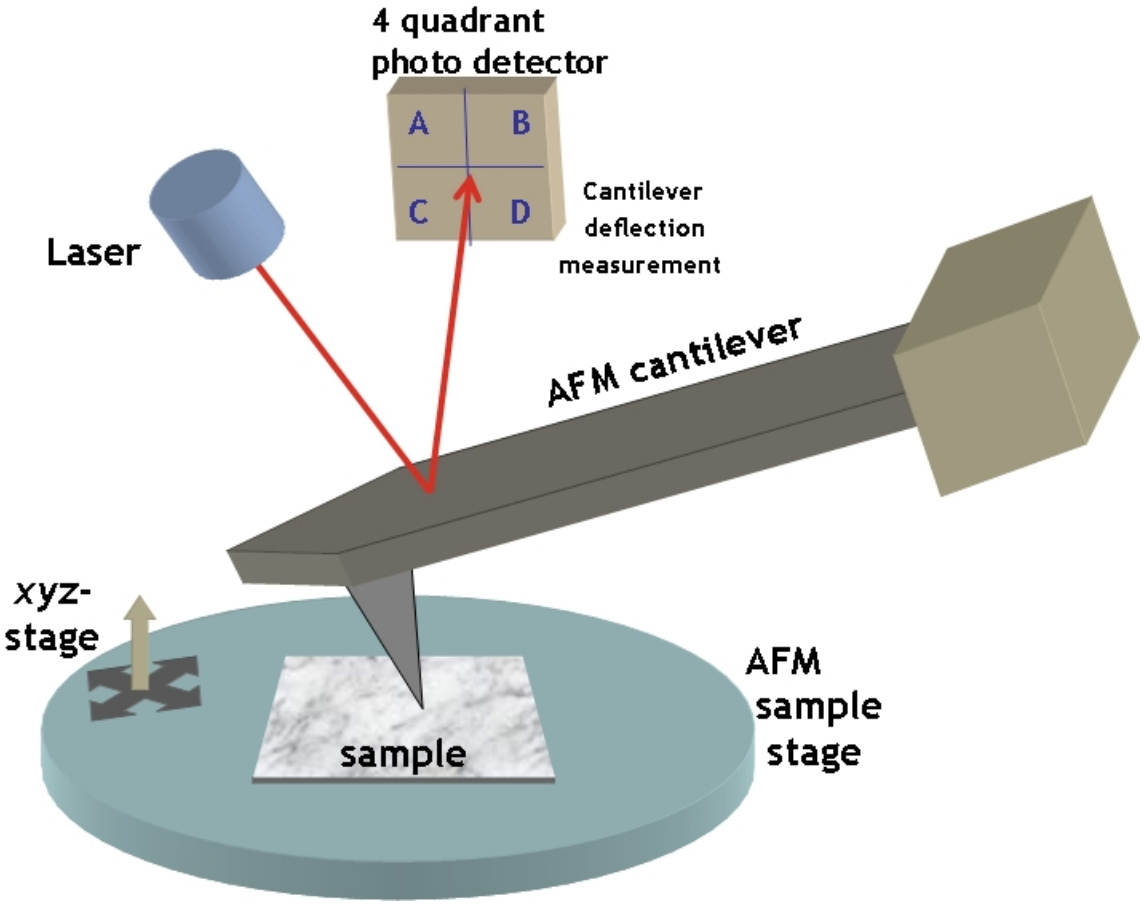
These subfields seek to anticipate what inventions nanotechnology might yield, or attempt to propose an agenda along which inquiry might progress. These often take a big-picture view of nanotechnology, with more emphasis on its societal implications than the details of how such inventions could actually be created.

- Molecular nanotechnology is a proposed approach which involves manipulating single molecules in finely controlled, deterministic ways. This is more theoretical than the other subfields and is beyond current capabilities.
- Nanorobotics centers on self-sufficient machines of some functionality operating at the nanoscale. There are hopes for applying nanorobots in medicine, but it may not be easy to do such a thing because of several drawbacks of such devices. Nevertheless, progress on innovative materials and methodologies has been demonstrated with some patents granted about new nanomanufacturing devices for future commercial applications, which also progressively helps in the development towards nanorobots with the use of embedded nanobioelectronics concepts.
- Productive nanosystems are "systems of nanosystems" which will be complex nanosystems that produce atomically precise parts for other nanosystems, not necessarily using novel nanoscale-emergent properties, but well-understood fundamentals of manufacturing. Because of the discrete (i.e. atomic) nature of matter and the possibility of exponential growth, this stage is seen as the basis of another industrial revolution. Mihail Roco, one of the architects of the USA's National Nanotechnology Initiative, has proposed four states of nanotechnology that seem to parallel the technical progress of the Industrial Revolution,

progressing from passive nanostructures to active nanodevices to complex nanomachines and ultimately to productive nanosystems.

- Programmable matter seeks to design materials whose properties can be easily, reversibly and externally controlled though a fusion of information science and materials science.
- Due to the popularity and media exposure of the term nanotechnology, the words picotechnology and femtotechnology have been coined in analogy to it, although these are only used rarely and informally.

Tools and techniques



Typical AFM setup. A microfabricated cantilever with a sharp tip is deflected by features on a sample surface, much like in a phonograph but on a much smaller scale. A laser beam reflects off the backside of the cantilever into a set of photodetectors, allowing the deflection to be measured and assembled into an image of the surface.

There are several important modern developments. The atomic force microscope (AFM) and the Scanning Tunneling Microscope (STM) are two early versions of scanning probes that launched nanotechnology. There are other types of scanning probe microscopy, all flowing from the ideas of the scanning confocal microscope developed by Marvin Minsky in 1961 and the scanning acoustic microscope (SAM) developed by

Calvin Quate and coworkers in the 1970s, that made it possible to see structures at the nanoscale. The tip of a scanning probe can also be used to manipulate nanostructures (a process called positional assembly). Feature-oriented scanning-positioning methodology suggested by Rostislav Lapshin appears to be a promising way to implement these nanomanipulations in automatic mode. However, this is still a slow process because of low scanning velocity of the microscope. Various techniques of nanolithography such as optical lithography, X-ray lithography dip pen nanolithography, electron beam lithography or nanoimprint lithography were also developed. Lithography is a top-down fabrication technique where a bulk material is reduced in size to nanoscale pattern.

Another group of nanotechnological techniques include those used for fabrication of nanowires, those used in semiconductor fabrication such as deep ultraviolet lithography, electron beam lithography, focused ion beam machining, nanoimprint lithography, atomic layer deposition, and molecular vapor deposition, and further including molecular self-assembly techniques such as those employing di-block copolymers. However, all of these techniques preceded the nanotech era, and are extensions in the development of scientific advancements rather than techniques which were devised with the sole purpose of creating nanotechnology and which were results of nanotechnology research.

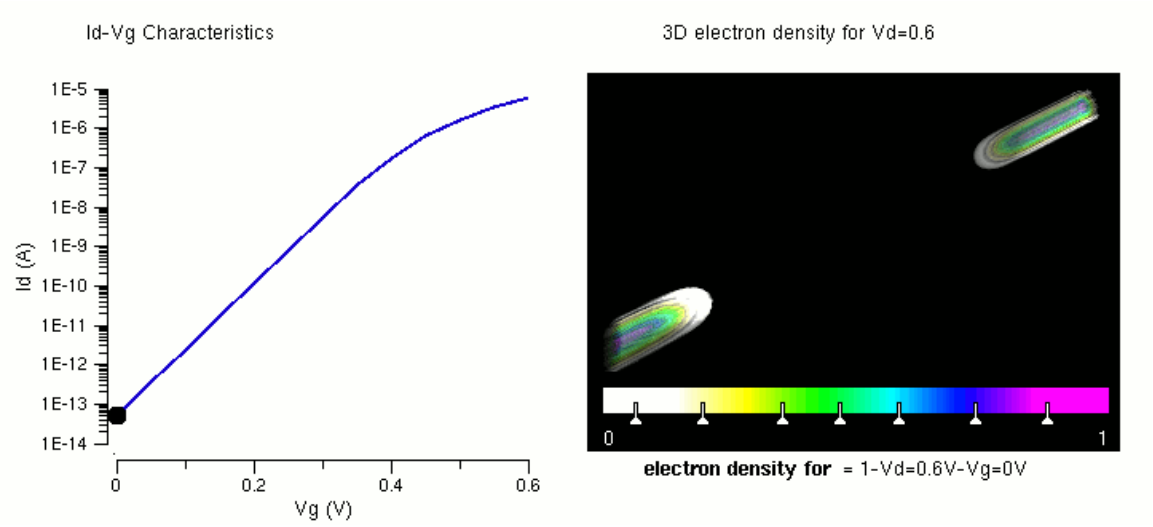
The top-down approach anticipates nanodevices that must be built piece by piece in stages, much as manufactured items are made. Scanning probe microscopy is an important technique both for characterization and synthesis of nanomaterials. Atomic force microscopes and scanning tunneling microscopes can be used to look at surfaces and to move atoms around. By designing different tips for these microscopes, they can be used for carving out structures on surfaces and to help guide self-assembling structures. By using, for example, feature-oriented scanning-positioning approach, atoms can be moved around on a surface with scanning probe microscopy techniques. At present, it is expensive and time-consuming for mass production but very suitable for laboratory experimentation.

In contrast, bottom-up techniques build or grow larger structures atom by atom or molecule by molecule. These techniques include chemical synthesis, self-assembly and positional assembly. Dual polarisation interferometry is one tool suitable for characterisation of self assembled thin films. Another variation of the bottom-up approach is molecular beam epitaxy or MBE. Researchers at Bell Telephone Laboratories like John R. Arthur. Alfred Y. Cho, and Art C. Gossard developed and implemented MBE as a research tool in the late 1960s and 1970s. Samples made by MBE were key to the discovery of the fractional quantum Hall effect for which the 1998 Nobel Prize in Physics was awarded. MBE allows scientists to lay down atomically precise layers of atoms and, in the process, build up complex structures. Important for research on semiconductors, MBE is also widely used to make samples and devices for the newly emerging field of spintronics.

However, new therapeutic products, based on responsive nanomaterials, such as the ultradeformable, stress-sensitive Transfersome vesicles, are under development and already approved for human use in some countries.

Applications

As of August 21, 2008, the Project on Emerging Nanotechnologies estimates that over 800 manufacturer-identified nanotech products are publicly available, with new ones hitting the market at a pace of 3–4 per week. The project lists all of the products in a publicly accessible online. Most applications are limited to the use of "first generation" passive nanomaterials which includes titanium dioxide in sunscreen, cosmetics and some food products; Carbon allotropes used to produce gecko tape; silver in food packaging, clothing, disinfectants and household appliances; zinc oxide in sunscreens and cosmetics, surface coatings, paints and outdoor furniture varnishes; and cerium oxide as a fuel catalyst.



One of the major applications of nanotechnology is in the area of nanoelectronics with MOSFET's being made of small nanowires ~10 nm in length. Here is a simulation of such a nanowire.

The National Science Foundation (a major distributor for nanotechnology research in the United States) funded researcher David Berube to study the field of nanotechnology. His findings are published in the monograph Nano-Hype: The Truth Behind the Nanotechnology Buzz. This study concludes that much of what is sold as "nanotechnology" is in fact a recasting of straightforward materials science, which is leading to a "nanotech industry built solely on selling nanotubes, nanowires, and the like" which will "end up with a few suppliers selling low margin products in huge volumes." Further applications which require actual manipulation or arrangement of nanoscale components await further research. Though technologies branded with the term 'nano' are sometimes little related to and fall far short of the most ambitious and transformative technological goals of the sort in molecular manufacturing proposals, the term still connotes such ideas. According to Berube, there may be a danger that a "nano bubble" will form, or is forming already, from the use of the term by scientists and entrepreneurs to garner funding, regardless of interest in the transformative possibilities of more ambitious and far-sighted work.

Implications

Because of the far-ranging claims that have been made about potential applications of nanotechnology, a number of serious concerns have been raised about what effects these will have on our society if realized, and what action if any is appropriate to mitigate these risks.

There are possible dangers that arise with the development of nanotechnology. The Center for Responsible Nanotechnology suggests that new developments could result, among other things, in untraceable weapons of mass destruction, networked cameras for use by the government, and weapons developments fast enough to destabilize arms races ("Nanotechnology Basics").

One area of concern is the effect that industrial-scale manufacturing and use of nanomaterials would have on human health and the environment, as suggested by nanotoxicology research. Groups such as the Center for Responsible Nanotechnology have advocated that nanotechnology should be specially regulated by governments for these reasons. Others counter that overregulation would stifle scientific research and the development of innovations which could greatly benefit mankind.

Other experts, including director of the Woodrow Wilson Center's Project on Emerging Nanotechnologies David Rejeski, have testified that successful commercialization depends on adequate oversight, risk research strategy, and public engagement. Berkeley, California is currently the only city in the United States to regulate nanotechnology; Cambridge, Massachusetts in 2008 considered enacting a similar law, but ultimately rejected this.

Health and environmental concerns

Some of the recently developed nanoparticle products may have unintended consequences. Researchers have discovered that silver nanoparticles used in socks only to reduce foot odor are being released in the wash with possible negative consequences. Silver nanoparticles, which are bacteriostatic, may then destroy beneficial bacteria which are important for breaking down organic matter in waste treatment plants or farms.

A study at the University of Rochester found that when rats breathed in nanoparticles, the particles settled in the brain and lungs, which led to significant increases in biomarkers for inflammation and stress response. A study in China indicated that nanoparticles induce skin aging through oxidative stress in hairless mice.

A two-year study at UCLA's School of Public Health found lab mice consuming nano-titanium dioxide showed DNA and chromosome damage to a degree "linked to all the big killers of man, namely cancer, heart disease, neurological disease and aging".

A major study published more recently in Nature Nanotechnology suggests some forms of carbon nanotubes – a poster child for the “nanotechnology revolution” – could be as

harmful as asbestos if inhaled in sufficient quantities. Anthony Seaton of the Institute of Occupational Medicine in Edinburgh, Scotland, who contributed to the article on carbon nanotubes said "We know that some of them probably have the potential to cause mesothelioma. So those sorts of materials need to be handled very carefully." In the absence of specific nano-regulation forthcoming from governments, Paull and Lyons (2008) have called for an exclusion of engineered nanoparticles from organic food. A newspaper article reports that workers in a paint factory developed serious lung disease and nanoparticles were found in their lungs.

Regulation

Calls for tighter regulation of nanotechnology have occurred alongside a growing debate related to the human health and safety risks associated with nanotechnology. Furthermore, there is significant debate about who is responsible for the regulation of nanotechnology. While some non-nanotechnology specific regulatory agencies currently cover some products and processes (to varying degrees) – by "bolting on" nanotechnology to existing regulations – there are clear gaps in these regimes. In "Nanotechnology Oversight: An Agenda for the Next Administration," former EPA deputy administrator J. Clarence (Terry) Davies lays out a clear regulatory roadmap for the next presidential administration and describes the immediate and longer term steps necessary to deal with the current shortcomings of nanotechnology oversight.

Stakeholders concerned by the lack of a regulatory framework to assess and control risks associated with the release of nanoparticles and nanotubes have drawn parallels with bovine spongiform encephalopathy ('mad cow's disease), thalidomide, genetically modified food, nuclear energy, reproductive technologies, biotechnology, and asbestosis. Dr. Andrew Maynard, chief science advisor to the Woodrow Wilson Center's Project on Emerging Nanotechnologies, concludes (among others) that there is insufficient funding for human health and safety research, and as a result there is currently limited understanding of the human health and safety risks associated with nanotechnology. As a result, some academics have called for stricter application of the precautionary principle, with delayed marketing approval, enhanced labelling and additional safety data development requirements in relation to certain forms of nanotechnology.

The Royal Society report identified a risk of nanoparticles or nanotubes being released during disposal, destruction and recycling, and recommended that "manufacturers of products that fall under extended producer responsibility regimes such as end-of-life regulations publish procedures outlining how these materials will be managed to minimize possible human and environmental exposure" (p.xiii). Reflecting the challenges for ensuring responsible life cycle regulation, the Institute for Food and Agricultural Standards has proposed standards for nanotechnology research and development should be integrated across consumer, worker and environmental standards. They also propose that NGOs and other citizen groups play a meaningful role in the development of these standards.

Chapter 8

Optics

Optics includes study of dispersion of light

Optics is the branch of physics which involves the behavior and properties of light, including its interactions with matter and the construction of instruments that use or detect it. Optics usually describes the behavior of visible, ultraviolet, and infrared light. Because light is an electromagnetic wave, other forms of electromagnetic radiation such as X-rays, microwaves, and radio waves exhibit similar properties.

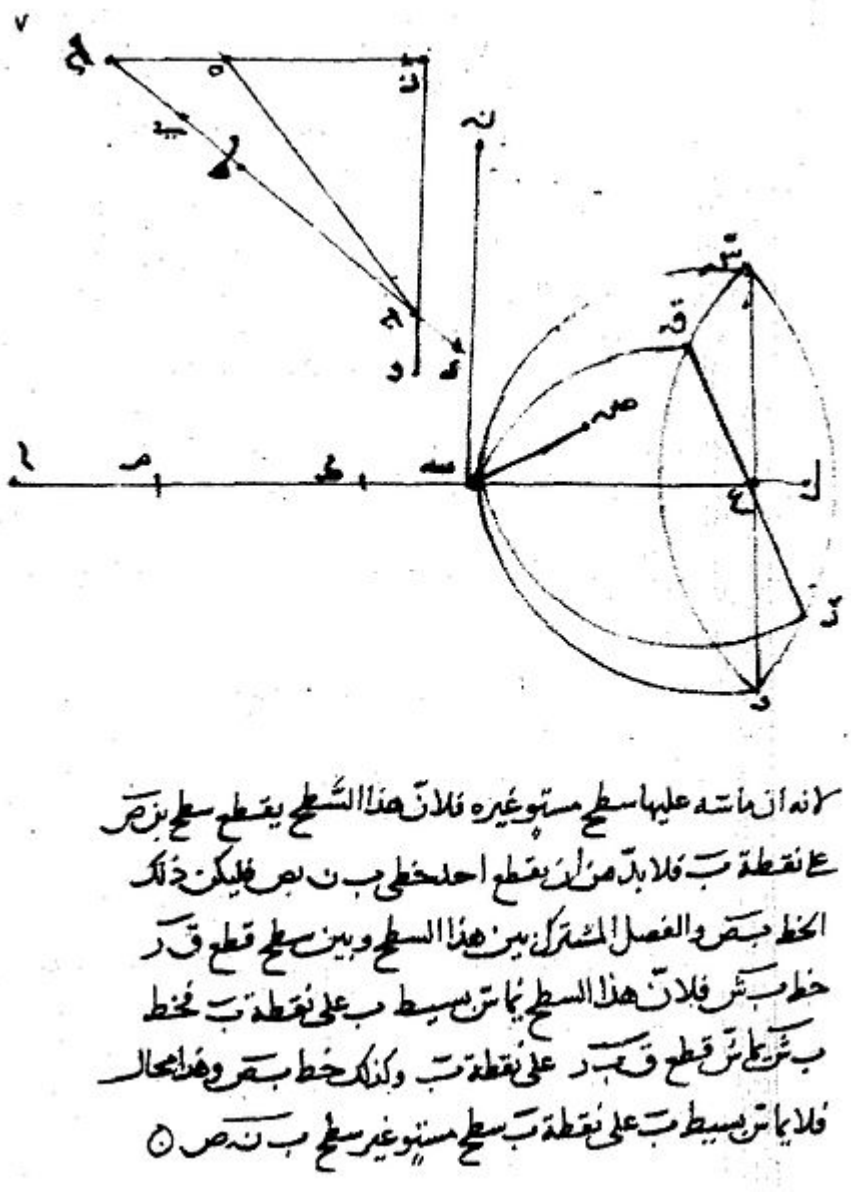
Most optical phenomena can be accounted for using the classical electromagnetic description of light. Complete electromagnetic descriptions of light are, however, often difficult to apply in practice. Practical optics is usually done using simplified models. The most common of these, geometric optics, treats light as a collection of rays that travel in straight lines and bend when they pass through or reflect from surfaces. Physical optics is a more comprehensive model of light, which includes wave effects such as diffraction and interference that cannot be accounted for in geometric optics. Historically, the ray-based model of light was developed first, followed by the wave model of light. Progress in electromagnetic theory in the 19th century led to the discovery that light waves were in fact electromagnetic radiation.

Some phenomena depend on the fact that light has both wave-like and particle-like properties. Explanation of these effects requires quantum mechanics. When considering light's particle-like properties, the light is modeled as a collection of particles called "photons". Quantum optics deals with the application of quantum mechanics to optical systems.

Optical science is relevant to and studied in many related disciplines including astronomy, various engineering fields, photography, and medicine (particularly ophthalmology and optometry). Practical applications of optics are found in a variety of technologies and everyday objects, including mirrors, lenses, telescopes, microscopes, lasers, and fiber optics.

History

Optics began with the development of lenses by the ancient Egyptians and Mesopotamians. The earliest known lenses were made from polished crystal, often quartz, and have been dated as early as 700 BC for Assyrian lenses such as the Layard/Nimrud lens. The ancient Romans and Greeks filled glass spheres with water to make lenses. These practical developments were followed by the development of theories of light and vision by ancient Greek and Indian philosophers, and the development of geometrical optics in the Greco-Roman world. The word *optics* comes from the ancient Greek word *ὀπτική*, meaning *appearance* or *look*. Plato first articulated emission theory, the idea that visual perception is accomplished by rays emitted by the eyes. He also commented on the parity reversal of mirrors in *Timaeus*. Some hundred years later, Euclid wrote a treatise entitled *Optics* wherein he described the mathematical rules of perspective and describes the effects of refraction qualitatively. Ptolemy, in his treatise *Optics*, summarizes much of Euclid and goes on to describe a way to measure the angle of refraction, though he failed to notice the empirical relationship between it and the angle of incidence.



Reproduction of a page of Ibn Sahl's manuscript showing his knowledge of the law of refraction, now known as Snell's law.

During the Middle Ages, Greek ideas about optics were resurrected and extended by writers in the Muslim world. One of the earliest of these was Al-Kindi (c. 801–73). In 984, the Persian mathematician Ibn Sahl wrote the treatise "On burning mirrors and lenses", correctly describing a law of refraction equivalent to Snell's law. He used this law to compute optimum shapes for lenses and curved mirrors. In the early 11th century, Alhazen (Ibn al-Haytham) wrote his *Book of Optics*, which documented the then-current understanding of vision.

In the 13th century, Roger Bacon used parts of glass spheres as magnifying glasses, and discovered that light reflects from objects rather than being released from them. In Italy, around 1284, Salvino D'Armate invented the first wearable eyeglasses.

The earliest known telescopes were refracting telescopes, a type which relies entirely on lenses for magnification. The first rudimentary telescopes were developed independently in the 1570s and 1580s by Leonard Digges, and Giambattista della Porta. Their development in the Netherlands in 1608 was by three individuals: Hans Lippershey and Zacharias Janssen, who were spectacle makers in Middelburg, and Jacob Metius of Alkmaar. In Italy, Galileo greatly improved upon these designs the following year. In 1668, Isaac Newton constructed the first practical reflecting telescope, which bears his name, the Newtonian reflector.

The first microscope was made around 1595, also in Middelburg. Three different eyeglass makers have been given credit for the invention: Lippershey, Janssen, and his father, Hans. The coining of the name "microscope" has been credited to Giovanni Faber, who gave that name to Galileo's compound microscope in 1625.

Optical theory progressed in the mid-17th century with treatises written by philosopher René Descartes, which explained a variety of optical phenomena including reflection and refraction by assuming that light was emitted by objects which produced it. This differed substantively from the ancient Greek emission theory. In the late 1660s and early 1670s, Newton expanded Descartes' ideas into a corpuscle theory of light, famously showing that white light, instead of being a unique color, was really a composite of different colors that can be separated into a spectrum with a prism. In 1690, Christian Huygens proposed a wave theory for light based on suggestions that had been made by Robert Hooke in 1664. Hooke himself publicly criticized Newton's theories of light and the feud between the two lasted until Hooke's death. In 1704, Newton published *Opticks* and, at the time, partly because of his success in other areas of physics, he was generally considered to be the victor in the debate over the nature of light.

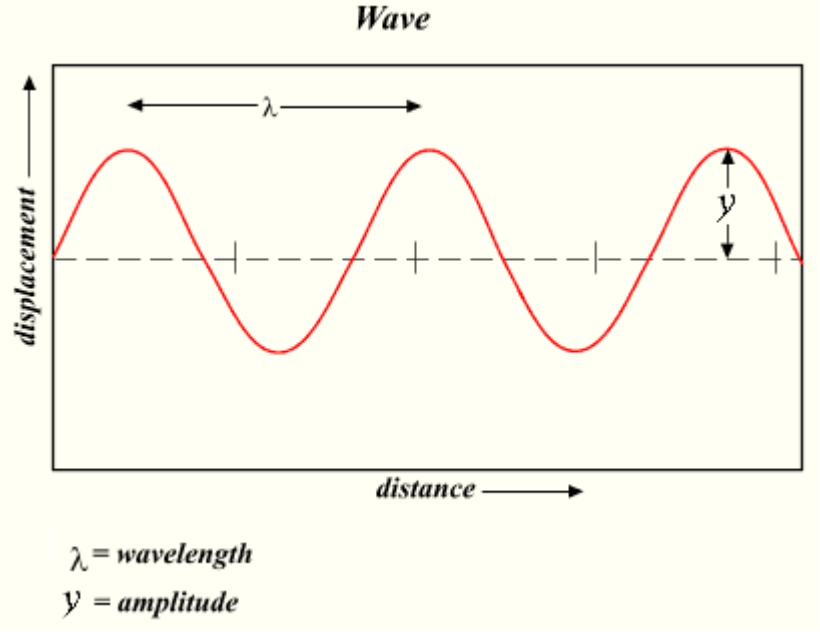
Newtonian optics was generally accepted until the early 19th century when Thomas Young and Augustin-Jean Fresnel conducted experiments on the interference of light that firmly established light's wave nature. Young's famous double slit experiment showed that light followed the law of superposition, which is a wave-like property not predicted by Newton's corpuscle theory. This work led to a theory of diffraction for light and opened an entire area of study in physical optics. Wave optics was successfully unified with electromagnetic theory by James Clerk Maxwell in the 1860s.

The next development in optical theory came in 1899 when Max Planck correctly modeled blackbody radiation by assuming that the exchange of energy between light and matter only occurred in discrete amounts he called *quanta*. In 1905, Albert Einstein published the theory of the photoelectric effect that firmly established the quantization of light itself. In 1913, Niels Bohr showed that atoms could only emit discrete amounts of energy, thus explaining the discrete lines seen in emission and absorption spectra. The understanding of the interaction between light and matter, which followed from these

developments, not only formed the basis of quantum optics but also was crucial for the development of quantum mechanics as a whole. The ultimate culmination was the theory of quantum electrodynamics, which explains all optics and electromagnetic processes in general as being the result of the exchange of real and virtual photons.

Quantum optics gained practical importance with the invention of the maser in 1953 and the laser in 1960. Following the work of Paul Dirac in quantum field theory, George Sudarshan, Roy J. Glauber, and Leonard Mandel applied quantum theory to the electromagnetic field in the 1950s and 1960s to gain a more detailed understanding of photodetection and the statistics of light.

Classical optics



Light propagates through space as a wave with amplitude, wavelength, frequency, and speed that depend on how it was emitted and on the medium through which it travels.

In pre-quantum-mechanical optics, light is an electromagnetic wave composed of oscillating electric and magnetic fields. These fields continually generate each other, as the wave propagates through space and oscillates in time.

The frequency of a light wave is determined by the period of the oscillations. The frequency does not normally change as the wave travels through different materials ("media"), but the speed of the wave depends on the medium. The speed, frequency, and wavelength of a wave are related by the formula

$$v = \lambda f,$$

where v is the speed, λ is the wavelength and f is the frequency. Because the frequency is fixed, a change in the wave's speed produces a change in its wavelength.

The speed of light in a medium is typically characterized by the index of refraction, n, which is the ratio of the speed of light in vacuum, c, to the speed in the medium:

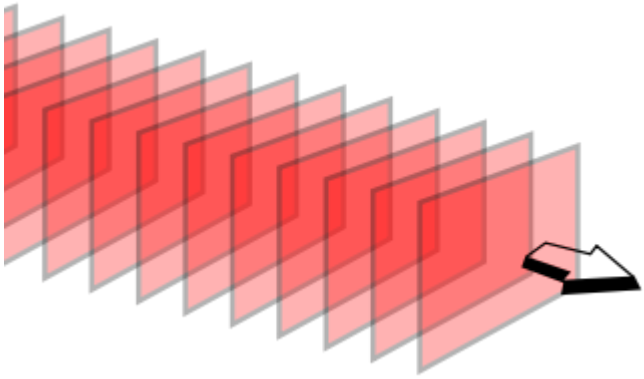
$$n = c / v.$$

The speed of light in vacuum is a constant, which is exactly 299,792,458 metres per second. Thus, a light ray with a wavelength of λ in a vacuum will have a wavelength of λ / n in a material with index of refraction n.

The amplitude of the light wave is related to the intensity of the light, which is related to the energy stored in the wave's electric and magnetic fields.

Traditional optics is divided into two main branches: geometrical optics and physical optics.

Geometrical optics



As a light wave travels through space, it oscillates in amplitude. In this image, each maximum amplitude crest is marked with a plane to illustrate the wavefront. The ray is the arrow perpendicular to these parallel surfaces.

Geometrical optics, or *ray optics*, describes light propagation in terms of "rays". The "ray" in geometric optics is an abstraction, or "instrument", that can be used to predict the path of light. A light ray is a ray that is perpendicular to the light's wavefronts (and therefore collinear with the wave vector). Light rays bend at the interface between two dissimilar media and may be curved in a medium in which the refractive index changes. Geometrical optics provides rules for propagating these rays through an optical system, which indicates how the actual wavefront will propagate. This is a significant simplification of optics that fails to account for optical effects such as diffraction and polarization. It is a good approximation, however, when the wavelength is very small compared with the size of structures with which the light interacts. Geometric optics can be used to describe the geometrical aspects of imaging, including optical aberrations.

A slightly more rigorous definition of a light ray follows from Fermat's principle which states that *the path taken between two points by a ray of light is the path that can be traversed in the least time.*

Approximations

Geometrical optics is often simplified by making the paraxial approximation, or "small angle approximation." The mathematical behavior then becomes linear, allowing optical components and systems to be described by simple matrices. This leads to the techniques of Gaussian optics and *paraxial ray tracing*, which are used to find basic properties of optical systems, such as approximate image and object positions and magnifications.

Reflections

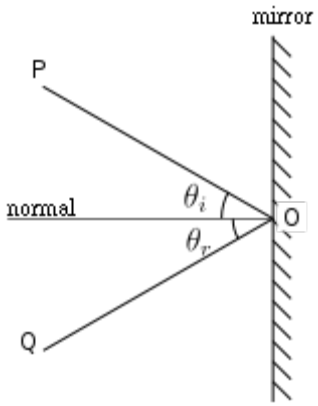


Diagram of specular reflection

Reflections can be divided into two types: specular reflection and diffuse reflection. Specular reflection describes the gloss of surfaces such as mirrors, which reflect light in a simple, predictable way. This allows for production of reflected images that can be associated with an actual (real) or extrapolated (virtual) location in space. Diffuse reflection describes opaque, non limpid materials, such as paper or rock. The reflections from these surfaces can only be described statistically, with the exact distribution of the reflected light depending on the microscopic structure of the material. Many diffuse reflectors are described or can be approximated by Lambert's cosine law, which describes surfaces that have equal luminance when viewed from any angle. Glossy surfaces can give both specular and diffuse reflection.

In specular reflection, the direction of the reflected ray is determined by the angle the incident ray makes with the surface normal, a line perpendicular to the surface at the point where the ray hits. The incident and reflected rays and the normal lie in a single plane, and the angle between the reflected ray and the surface normal is the same as that between the incident ray and the normal. This is known as the Law of Reflection.

For flat mirrors, the law of reflection implies that images of objects are upright and the same distance behind the mirror as the objects are in front of the mirror. The image size is

the same as the object size. (The magnification of a flat mirror is unity.) The law also implies that mirror images are parity inverted, which we perceive as a left-right inversion. Images formed from reflection in two (or any even number of) mirrors are not parity inverted. Corner reflectors retroreflect light, producing reflected rays that travel back in the direction from which the incident rays came.

Mirrors with curved surfaces can be modeled by ray-tracing and using the law of reflection at each point on the surface. For mirrors with parabolic surfaces, parallel rays incident on the mirror produce reflected rays that converge at a common focus. Other curved surfaces may also focus light, but with aberrations due to the diverging shape causing the focus to be smeared out in space. In particular, spherical mirrors exhibit spherical aberration. Curved mirrors can form images with magnification greater than or less than one, and the magnification can be negative, indicating that the image is inverted. An upright image formed by reflection in a mirror is always virtual, while an inverted image is real and can be projected onto a screen.

Refractions

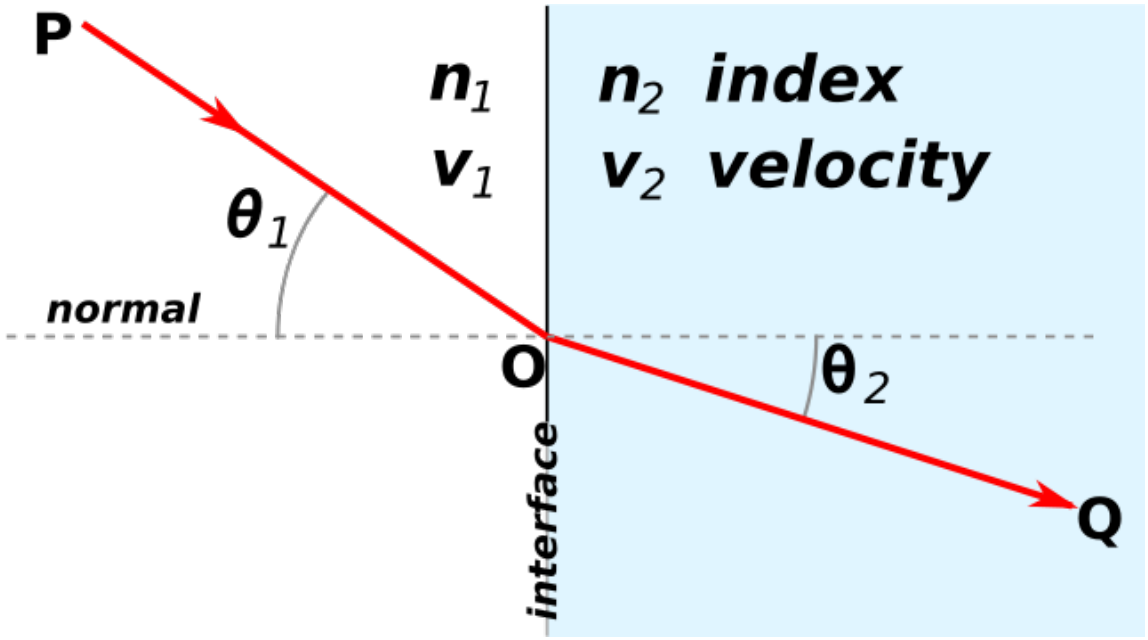


Illustration of Snell's Law for the case $n_1 < n_2$, such as air/water interface

Refraction occurs when light travels through an area of space that has a changing index of refraction; this principle allows for lenses and the focusing of light. The simplest case of refraction occurs when there is an interface between a uniform medium with index of refraction n_1 and another medium with index of refraction n_2. In such situations, Snell's Law describes the resulting deflection of the light ray:

$$n_1 \sin\theta_1 = n_2 \sin\theta_2$$

where θ_1 and θ_2 are the angles between the normal (to the interface) and the incident and refracted waves, respectively. This phenomenon is also associated with a changing speed of light as seen from the definition of index of refraction provided above which implies:

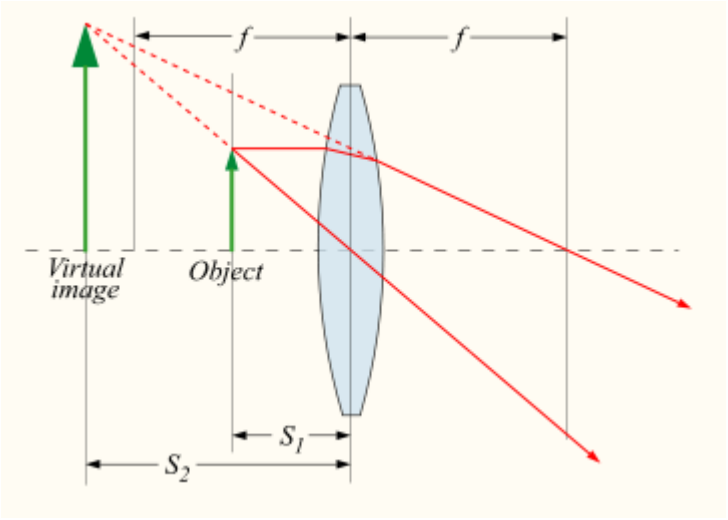
$$v_1 \sin\theta_2 = v_2 \sin\theta_1$$

where v_1 and v_2 are the wave velocities through the respective media.

Various consequences of Snell's Law include the fact that for light rays traveling from a material with a high index of refraction to a material with a low index of refraction, it is possible for the interaction with the interface to result in zero transmission. This phenomenon is called total internal reflection and allows for fiber optics technology. As light signals travel down a fiber optic cable, it undergoes total internal reflection allowing for essentially no light lost over the length of the cable. It is also possible to produce polarized light rays using a combination of reflection and refraction: When a refracted ray and the reflected ray form a right angle, the reflected ray has the property of "plane polarization". The angle of incidence required for such a scenario is known as Brewster's angle.

Snell's Law can be used to predict the deflection of light rays as they pass through "linear media" as long as the indexes of refraction and the geometry of the media are known. For example, the propagation of light through a prism results in the light ray being deflected depending on the shape and orientation of the prism. Additionally, since different frequencies of light have slightly different indexes of refraction in most materials, refraction can be used to produce dispersion spectra that appear as rainbows. The discovery of this phenomenon when passing light through a prism is famously attributed to Isaac Newton.

Some media have an index of refraction which varies gradually with position and, thus, light rays curve through the medium rather than travel in straight lines. This effect is what is responsible for mirages seen on hot days where the changing index of refraction of the air causes the light rays to bend creating the appearance of specular reflections in the distance (as if on the surface of a pool of water). Material that has a varying index of refraction is called a gradient-index (GRIN) material and has many useful properties used in modern optical scanning technologies including photocopiers and scanners. The phenomenon is studied in the field of gradient-index optics.

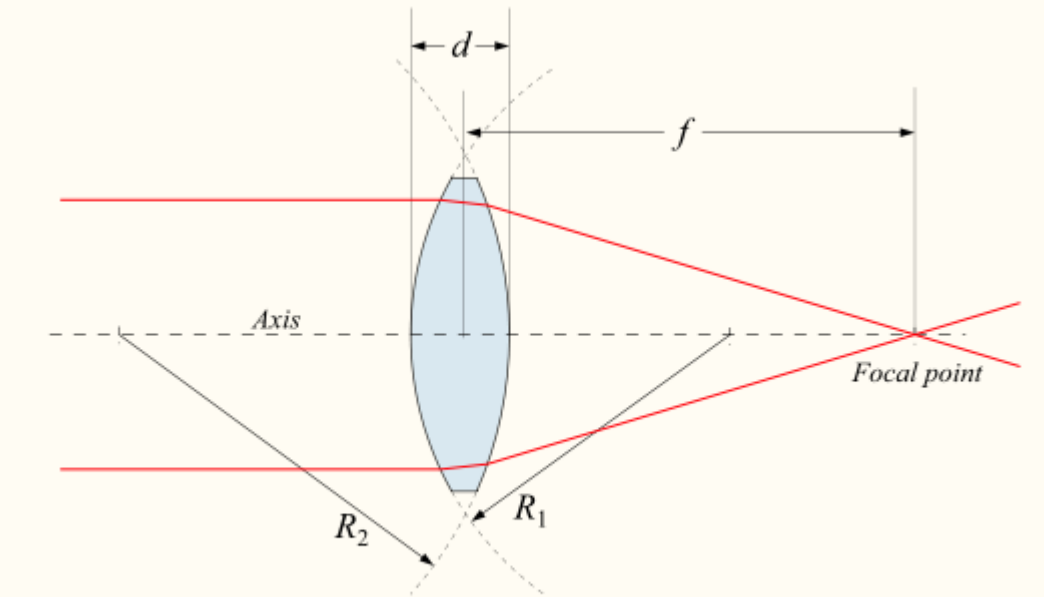


A ray tracing diagram for a converging lens.

A device which produces converging or diverging light rays due to refraction is known as a lens. Thin lenses produce focal points on either side that can be modeled using the lensmaker's equation. In general, two types of lenses exist: convex lenses, which cause parallel light rays to converge, and concave lenses, which cause parallel light rays to diverge. The detailed prediction of how images are produced by these lenses can be made using ray-tracing similar to curved mirrors. Similarly to curved mirrors, thin lenses follow a simple equation that determines the location of the images given a particular focal length (f) and object distance (S_1):

$$\frac{1}{S_1} + \frac{1}{S_2} = \frac{1}{f}$$

where S_2 is the distance associated with the image and is considered by convention to be negative if on the same side of the lens as the object and positive if on the opposite side of the lens. The focal length f is considered negative for concave lenses.



Positive (converging) lens

Incoming parallel rays are focused by a convex lens into an inverted real image one focal length from the lens, on the far side of the lens. Rays from an object at finite distance are focused further from the lens than the focal distance; the closer the object is to the lens, the further the image is from the lens. With concave lenses, incoming parallel rays diverge after going through the lens, in such a way that they seem to have originated at an upright virtual image one focal length from the lens, on the same side of the lens that the parallel rays are approaching on. Rays from an object at finite distance are associated with a virtual image that is closer to the lens than the focal length, and on the same side of the lens as the object. The closer the object is to the lens, the closer the virtual image is to the lens.

Likewise, the magnification of a lens is given by

$$M = -\frac{S_2}{S_1} = \frac{f}{f - S_1}$$

where the negative sign is given, by convention, to indicate an upright object for positive values and an inverted object for negative values. Similar to mirrors, upright images produced by single lenses are virtual while inverted images are real.

Lenses suffer from aberrations that distort images and focal points. These are due to both to geometrical imperfections and due to the changing index of refraction for different wavelengths of light (chromatic aberration).

Physical optics

Physical optics or wave optics builds on Huygens's principle, which states that every point on an advancing wavefront is the center of a new disturbance. When combined with the superposition principle, this explains how optical phenomena are manifested when there are multiple sources or obstructions that are spaced at distances similar to the wavelength of the light.

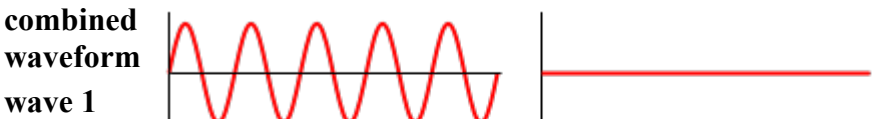
Complex models based on physical optics can account for the propagation of any wavefront through an optical system, including predicting the wavelength, amplitude, and phase of the wave. Additionally, all of the results from geometrical optics can be recovered using the techniques of Fourier optics which apply many of the same mathematical and analytical techniques used in acoustic engineering and signal processing.

Using numerical modeling on a computer, optical scientists can simulate the propagation of light and account for most diffraction, interference, and polarization effects. Such simulations typically still rely on approximations, however, so this is not a full electromagnetic wave theory model of the propagation of light. Such a full model is computationally demanding and is normally only used to solve small-scale problems that require extraordinary accuracy.

Gaussian beam propagation is a simple paraxial physical optics model for the propagation of coherent radiation such as laser beams. This technique partially accounts for diffraction, allowing accurate calculations of the rate at which a laser beam expands with distance, and the minimum size to which the beam can be focused. Gaussian beam propagation thus bridges the gap between geometric and physical optics.

Superposition and interference

In the absence of nonlinear effects, the superposition principle can be used to predict the shape of interacting waveforms through the simple addition of the disturbances. This interaction of waves to produce a resulting pattern is generally termed "interference" and can result in a variety of outcomes. If two waves of the same wavelength and frequency are *in phase*, both the wave crests and wave troughs align. This results in constructive interference and an increase in the amplitude of the wave, which for light is associated with a brightening of the waveform in that location. Alternatively, if the two waves of the same wavelength and frequency are out of phase, then the wave crests will align with wave troughs and vice-versa. This results in destructive interference and a decrease in the amplitude of the wave, which for light is associated with a dimming of the waveform at that location. See below for an illustration of this effect.



wave 2

Two waves in phase **Two waves 180° out of phase**

When oil or fuel is spilled, colorful patterns are formed by thin-film interference.

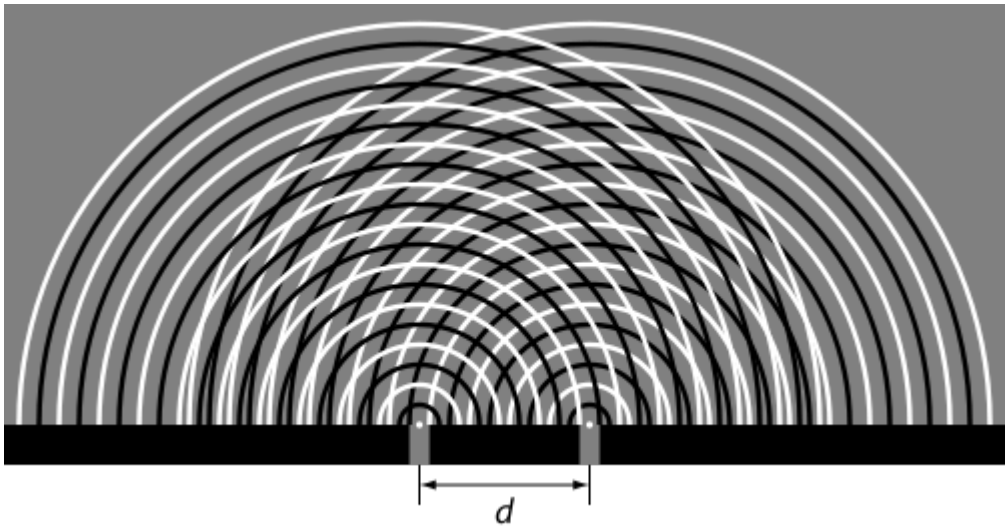
Since Huygens's principle states that every point of a wavefront is associated with the production of a new disturbance, it is possible for a wavefront to interfere with itself constructively or destructively at different locations producing bright and dark fringes in regular and predictable patterns. Interferometry is the science of measuring these patterns, usually as a means of making precise determinations of distances or angular resolutions. The Michelson interferometer was a famous instrument which used interference effects to accurately measure the speed of light.

The appearance of thin films and coatings is directly affected by interference effects. Antireflective coatings use destructive interference to reduce the reflectivity of the surfaces they coat, and can be used to minimize glare and unwanted reflections. The simplest case is a single layer with thickness one-fourth the wavelength of incident light.

The reflected wave from the top of the film and the reflected wave from the film/material interface are then exactly 180° out of phase, causing destructive interference. The waves are only exactly out of phase for one wavelength, which would typically be chosen to be near the center of the visible spectrum, around 550 nm. More complex designs using multiple layers can achieve low reflectivity over a broad band, or extremely low reflectivity at a single wavelength.

Constructive interference in thin films can create strong reflection of light in a range of wavelengths, which can be narrow or broad depending on the design of the coating. These films are used to make dielectric mirrors, interference filters, heat reflectors, and filters for color separation in color television cameras. This interference effect is also what causes the colorful rainbow patterns seen in oil slicks.

Diffraction and optical resolution



Diffraction on two slits separated by distance *d*. The bright fringes occur along lines where black lines intersect with black lines and white lines intersect with white lines. These fringes are separated by angle θ and are numbered as order *n*.

Diffraction is the process by which light interference is most commonly observed. The effect was first described in 1665 by Francesco Maria Grimaldi, who also coined the term from the Latin *diffringere*, 'to break into pieces'. Later that century, Robert Hooke and Isaac Newton also described phenomena now known to be diffraction in Newton's rings while James Gregory recorded his observations of diffraction patterns from bird feathers.

The first physical optics model of diffraction that relied on Huygens' Principle was developed in 1803 by Thomas Young in his accounts of the interference patterns of two closely spaced slits. Young showed that his results could only be explained if the two slits acted as two unique sources of waves rather than corpuscles. In 1815 and 1818, Augustin-Jean Fresnel firmly established the mathematics of how wave interference can account for diffraction.

The simplest physical models of diffraction use equations that describe the angular separation of light and dark fringes due to light of a particular wavelength (λ). In general, the equation takes the form

$$m\lambda = d\sin\theta$$

where d is the separation between two wavefront sources (in the case of Young's experiments, it was two slits), θ is the angular separation between the central fringe and the mth order fringe, where the central maximum is $m = 0$.

This equation is modified slightly to take into account a variety of situations such as diffraction through a single gap, diffraction through multiple slits, or diffraction through a diffraction grating that contains a large number of slits at equal spacing. More complicated models of diffraction require working with the mathematics of Fresnel or Fraunhofer diffraction.

X-ray diffraction makes use of the fact that atoms in a crystal have regular spacing at distances that are on the order of one angstrom. To see diffraction patterns, x-rays with similar wavelengths to that spacing are passed through the crystal. Since crystals are three-dimensional objects rather than two-dimensional gratings, the associated diffraction pattern varies in two directions according to Bragg reflection, with the associated bright spots occurring in unique patterns and d being twice the spacing between atoms.

Diffraction effects limit the ability for an optical detector to optically resolve separate light sources. In general, light that is passing through an aperture will experience diffraction and the best images that can be created (as described in diffraction-limited optics) appear as a central spot with surrounding bright rings, separated by dark nulls; this pattern is known as an Airy pattern, and the central bright lobe as an Airy disk. The size of such a disk is given by

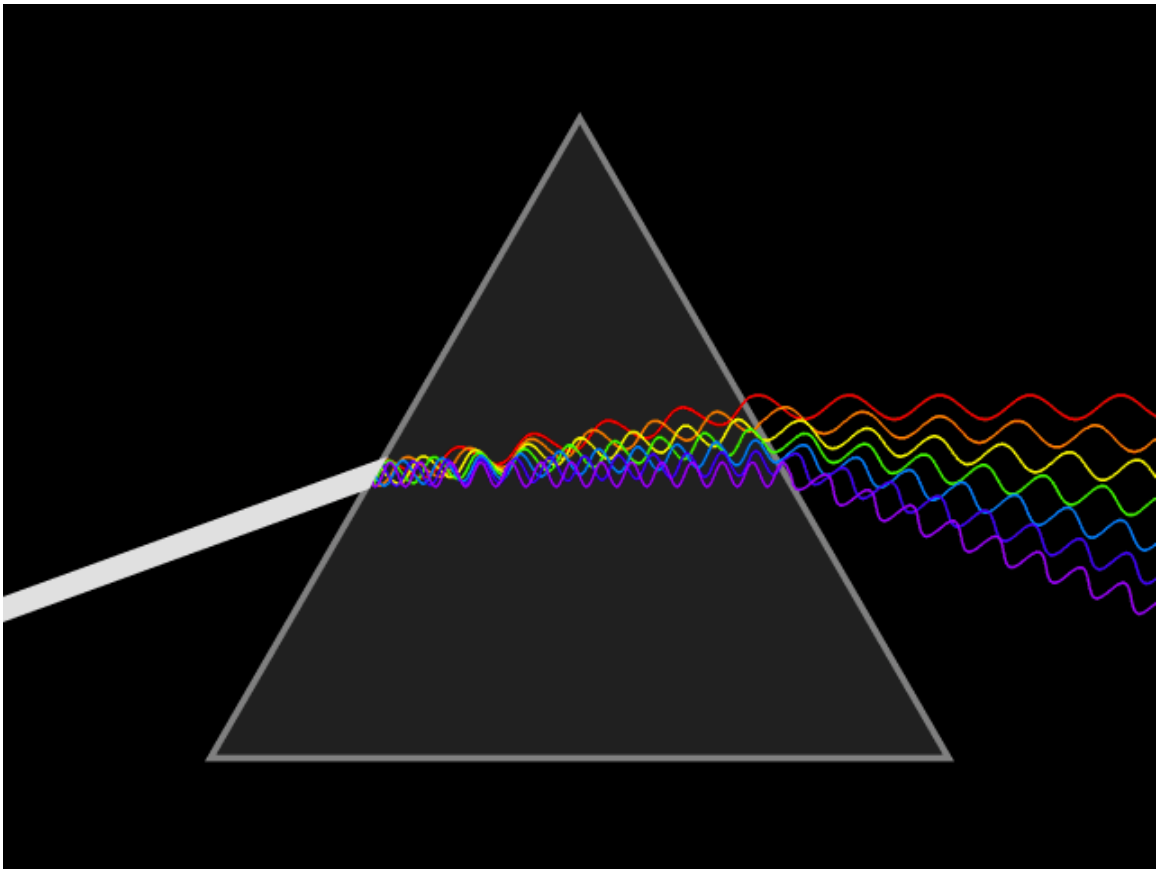
$$\sin\theta = 1.22\frac{\lambda}{D}$$

where θ is the angular resolution, λ is the wavelength of the light, and D is the diameter of the lens aperture. If the angular separation of the two points is significantly less than the Airy disk angular radius, then the two points cannot be resolved in the image, but if their angular separation is much greater than this, distinct images of the two points are formed and they can therefore be resolved. Rayleigh defined the somewhat arbitrary "Rayleigh criterion" that two points whose angular separation is equal to the Airy disk radius (measured to first null, that is, to the first place where no light is seen) can be considered to be resolved. It can be seen that the greater the diameter of the lens or its aperture, the finer the resolution. Interferometry, with its ability to mimic extremely large baseline apertures, allows for the greatest angular resolution possible.

For astronomical imaging, the atmosphere prevents optimal resolution from being achieved in the visible spectrum due to the atmospheric scattering and dispersion which

cause stars to twinkle. Astronomers refer to this effect as the quality of astronomical seeing. Techniques known as adaptive optics have been utilized to eliminate the atmospheric disruption of images and achieve results that approach the diffraction limit.

Dispersion and scattering



High frequency (blue) light is deflected the most, and low frequency (red) the least.

Refractive processes take place in the physical optics limit, where the wavelength of light is similar to other distances, as a kind of scattering. The simplest type of scattering is Thomson scattering which occurs when electromagnetic waves are deflected by single particles. In the limit of Thompson scattering, in which the wavelike nature of light is evident, light is dispersed independent of the frequency, in contrast to Compton scattering which is frequency-dependent and strictly a quantum mechanical process, involving the nature of light as particles. In a statistical sense, elastic scattering of light by numerous particles much smaller than the wavelength of the light is a process known as Rayleigh scattering while the similar process for scattering by particles that are similar or larger in wavelength is known as Mie scattering with the Tyndall effect being a commonly observed result. A small proportion of light scattering from atoms or molecules may undergo Raman scattering, wherein the frequency changes due to excitation of the atoms and molecules. Brillouin scattering occurs when the frequency of light changes due to local changes with time and movements of a dense material.

Dispersion occurs when different frequencies of light have different phase velocities, due either to material properties (*material dispersion*) or to the geometry of an optical waveguide (*waveguide dispersion*). The most familiar form of dispersion is a decrease in index of refraction with increasing wavelength, which is seen in most transparent materials. This is called "normal dispersion". It occurs in all dielectric materials, in wavelength ranges where the material does not absorb light. In wavelength ranges where a medium has significant absorption, the index of refraction can increase with wavelength. This is called "anomalous dispersion".

The separation of colors by a prism is an example of normal dispersion. At the surfaces of the prism, Snell's law predicts that light incident at an angle θ to the normal will be refracted at an angle arcsin(sin (θ) / *n*) . Thus, blue light, with its higher refractive index, is bent more strongly than red light, resulting in the well-known rainbow pattern.

Dispersion: two sinusoids propagating at different speeds make a moving interference pattern. The red dot moves with the phase velocity, and the green dots propagate with the group velocity. In this case, the phase velocity is twice the group velocity. The red dot overtakes two green dots, when moving from the left to the right of the figure. In effect, the individual waves (which travel with the phase velocity) escape from the wave packet (which travels with the group velocity).

Material dispersion is often characterized by the Abbe number, which gives a simple measure of dispersion based on the index of refraction at three specific wavelengths. Waveguide dispersion is dependent on the propagation constant. Both kinds of dispersion cause changes in the group characteristics of the wave, the features of the wave packet that change with the same frequency as the amplitude of the electromagnetic wave. "Group velocity dispersion" manifests as a spreading-out of the signal "envelope" of the radiation and can be quantified with a group dispersion delay parameter:

$$D = \frac{1}{v_g^2}\frac{dv_g}{d\lambda}$$

where v_g is the group velocity. For a uniform medium, the group velocity is

$$v_g = c\left(n - \lambda\frac{dn}{d\lambda}\right)^{-1}$$

where n is the index of refraction and c is the speed of light in a vacuum. This gives a simpler form for the dispersion delay parameter:

$$D = -\frac{\lambda}{c}\frac{d^2n}{d\lambda^2}.$$

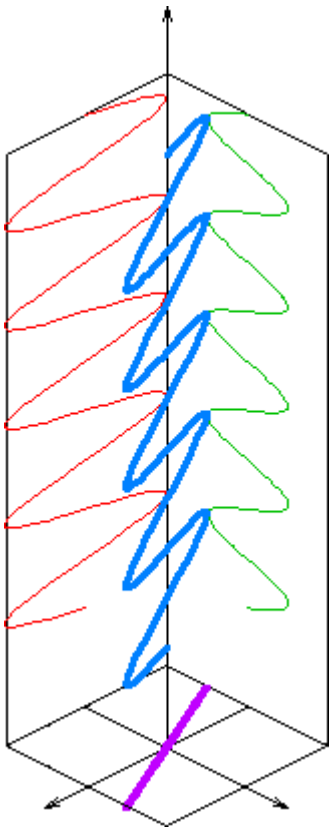
If *D* is less than zero, the medium is said to have *positive dispersion* or normal dispersion. If *D* is greater than zero, the medium has *negative dispersion*. If a light pulse is propagated through a normally dispersive medium, the result is the higher frequency components slow down more than the lower frequency components. The pulse therefore becomes *positively chirped*, or *up-chirped*, increasing in frequency with time. This causes the spectrum coming out of a prism to appear with red light the least refracted and blue/violet light the most refracted. Conversely, if a pulse travels through an anomalously (negatively) dispersive medium, high frequency components travel faster than the lower ones, and the pulse becomes *negatively chirped*, or *down-chirped*, decreasing in frequency with time.

The result of group velocity dispersion, whether negative or positive, is ultimately temporal spreading of the pulse. This makes dispersion management extremely important in optical communications systems based on optical fibers, since if dispersion is too high, a group of pulses representing information will each spread in time and merge together, making it impossible to extract the signal.

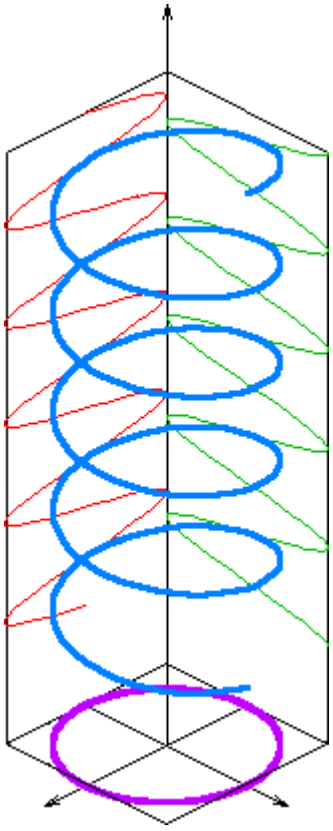
Polarization

Polarization is a general property of waves that describes the orientation of their oscillations. For transverse waves such as many electromagnetic waves, it describes the orientation of the oscillations in the plane perpendicular to the wave's direction of travel. The oscillations may be oriented in a single direction (linear polarization), or the oscillation direction may rotate as the wave travels (circular or elliptical polarization). Circularly polarized waves can rotate rightward or leftward in the direction of travel, and which of those two rotations is present in a wave is called the wave's chirality.

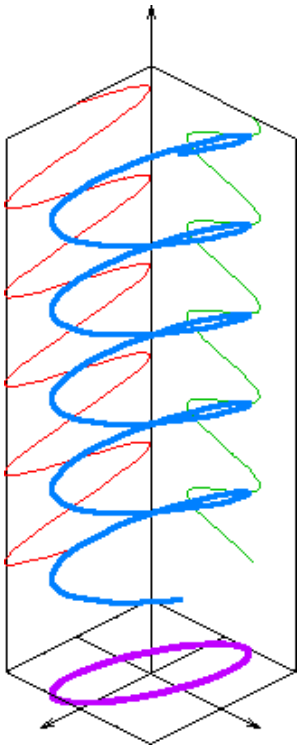
The typical way to consider polarization is to keep track of the orientation of the electric field vector as the electromagnetic wave propagates. The electric field vector of a plane wave may be arbitrarily divided into two perpendicular components labeled *x* and *y* (with **z** indicating the direction of travel). The shape traced out in the x-y plane by the electric field vector is a Lissajous figure that describes the *polarization state*. The following figures show some examples of the evolution of the electric field vector (blue), with time (the vertical axes), at a particular point in space, along with its *x* and *y* components (red/left and green/right), and the path traced by the vector in the plane (purple): The same evolution would occur when looking at the electric field at a particular time while evolving the point in space, along the direction opposite to propagation.



Linear



Circular



Elliptical polarization

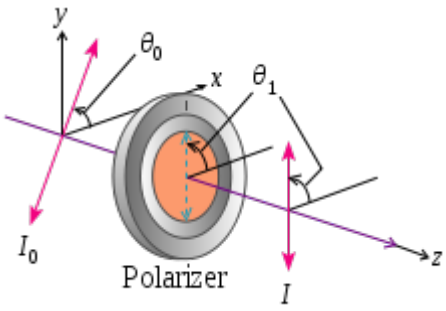
In the leftmost figure above, the x and y components of the light wave are in phase. In this case, the ratio of their strengths is constant, so the direction of the electric vector (the vector sum of these two components) is constant. Since the tip of the vector traces out a single line in the plane, this special case is called linear polarization. The direction of this line depends on the relative amplitudes of the two components.

In the middle figure, the two orthogonal components have the same amplitudes and are 90° out of phase. In this case, one component is zero when the other component is at maximum or minimum amplitude. There are two possible phase relationships that satisfy this requirement: the *x* component can be 90° ahead of the *y* component or it can be 90° behind the *y* component. In this special case, the electric vector traces out a circle in the plane, so this polarization is called circular polarization. The rotation direction in the circle depends on which of the two phase relationships exists and corresponds to *right-hand circular polarization* and *left-hand circular polarization*.

In all other cases, where the two components either do not have the same amplitudes and/or their phase difference is neither zero nor a multiple of 90°, the polarization is called elliptical polarization because the electric vector traces out an ellipse in the plane (the *polarization ellipse*). This is shown in the above figure on the right. Detailed

mathematics of polarization is done using Jones calculus and is characterized by the Stokes parameters.

Media that have different indexes of refraction for different polarization modes are called *birefringent*. Well known manifestations of this effect appear in optical wave plates/retarders (linear modes) and in Faraday rotation/optical rotation (circular modes). If the path length in the birefringent medium is sufficient, plane waves will exit the material with a significantly different propagation direction, due to refraction. For example, this is the case with macroscopic crystals of calcite, which present the viewer with two offset, orthogonally polarized images of whatever is viewed through them. It was this effect that provided the first discovery of polarization, by Erasmus Bartholinus in 1669. In addition, the phase shift, and thus the change in polarization state, is usually frequency dependent, which, in combination with dichroism, often gives rise to bright colors and rainbow-like effects. In mineralogy, such properties, known as pleochroism, are frequently exploited for the purpose of identifying minerals using polarization microscopes. Additionally, many plastics that are not normally birefringent will become so when subject to mechanical stress, a phenomenon which is the basis of photoelasticity. Non-birefringent methods, to rotate the linear polarization of light beams, include the use of prismatic polarization rotators which utilize total internal reflection in a prism set designed for efficient colinear transmission.



A polarizer changing the orientation of linearly polarized light. In this picture, $\theta_1 - \theta_0 = \theta_i$.

Media that reduce the amplitude of certain polarization modes are called *dichroic*. with devices that block nearly all of the radiation in one mode known as *polarizing filters* or simply "polarizers". Malus' law, which is named after Etienne-Louis Malus, says that when a perfect polarizer is placed in a linear polarized beam of light, the intensity, I, of the light that passes through is given by

$$I = I_0 \cos^2 \theta_i \quad ,$$

where

I_0 is the initial intensity,
and θ_i is the angle between the light's initial polarization direction and the axis of the polarizer.

A beam of unpolarized light can be thought of as containing a uniform mixture of linear polarizations at all possible angles. Since the average value of $\cos^2\theta$ is 1/2, the transmission coefficient becomes

$$\frac{I}{I_0} = \frac{1}{2}$$

In practice, some light is lost in the polarizer and the actual transmission of unpolarized light will be somewhat lower than this, around 38% for Polaroid-type polarizers but considerably higher (>49.9%) for some birefringent prism types.

In addition to birefringence and dichroism in extended media, polarization effects can also occur at the (reflective) interface between two materials of different refractive index. These effects are treated by the Fresnel equations. Part of the wave is transmitted and part is reflected, with the ratio depending on angle of incidence and the angle of refraction. In this way, physical optics recovers Brewster's angle.

The effects of a polarizing filter on the sky in a photograph. Left picture is taken without polarizer. For the right picture, filter was adjusted to eliminate certain polarizations of the scattered blue light from the sky.

Most sources of electromagnetic radiation contain a large number of atoms or molecules that emit light. The orientation of the electric fields produced by these emitters may not be correlated, in which case the light is said to be *unpolarized*. If there is partial correlation between the emitters, the light is *partially polarized*. If the polarization is consistent across the spectrum of the source, partially polarized light can be described as a superposition of a completely unpolarized component, and a completely polarized one. One may then describe the light in terms of the degree of polarization, and the parameters of the polarization ellipse.

Light reflected by shiny transparent materials is partly or fully polarized, except when the light is normal (perpendicular) to the surface. It was this effect that allowed the mathematician Etienne Louis Malus to make the measurements that allowed for his

development of the first mathematical models for polarized light. Polarization occurs when light is scattered in the atmosphere. The scattered light produces the brightness and color in clear skies. This partial polarization of scattered light can be taken advantage of using polarizing filters to darken the sky in photographs. Optical polarization is principally of importance in chemistry due to circular dichroism and optical rotation ("*circular birefringence*") exhibited by optically active (chiral) molecules.

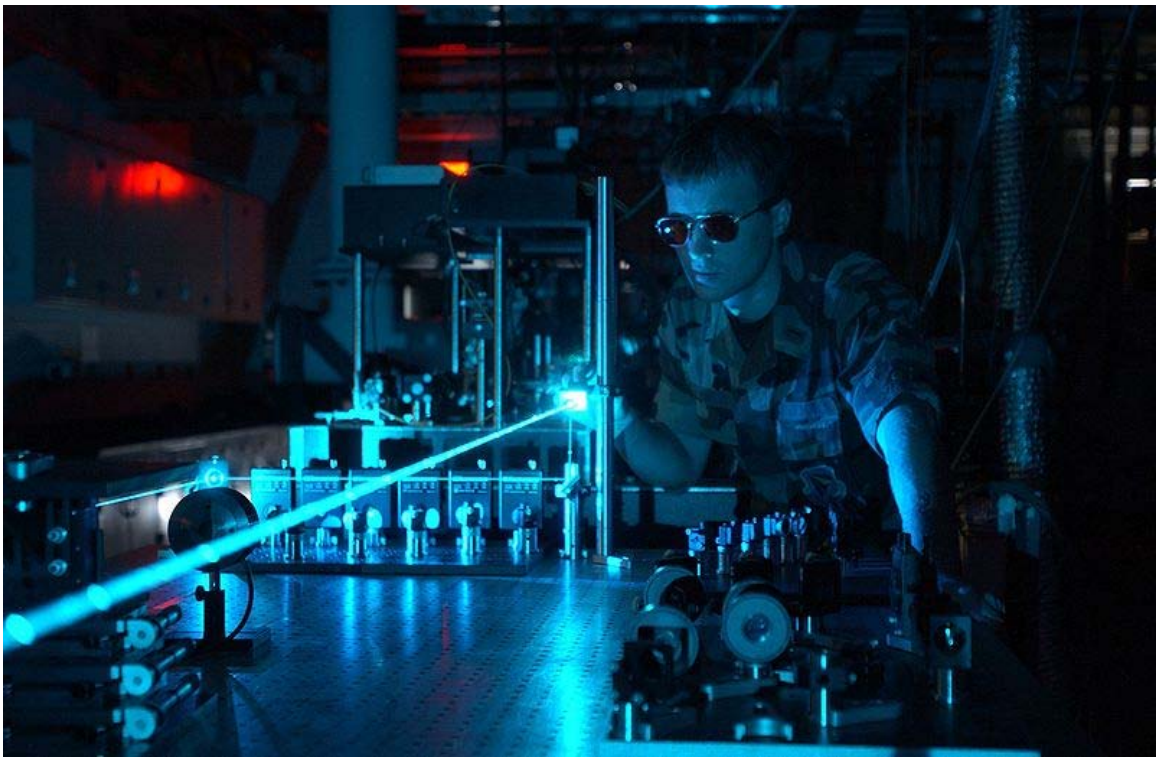
Modern optics

Modern optics encompasses the areas of optical science and engineering that became popular in the 20th century. These areas of optical science typically relate to the electromagnetic or quantum properties of light but do include other topics. A major subfield of modern optics, quantum optics, deals with specifically quantum mechanical properties of light. Quantum optics is not just theoretical; some modern devices, such as lasers, have principles of operation that depend on quantum mechanics. Light detectors, such as photomultipliers and channeltrons, respond to individual photons. Electronic image sensors, such as CCDs, exhibit shot noise corresponding to the statistics of individual photon events. Light-emitting diodes and photovoltaic cells, too, cannot be understood without quantum mechanics. In the study of these devices, quantum optics often overlaps with quantum electronics.

Specialty areas of optics research include the study of how light interacts with specific materials as in crystal optics and metamaterials. Other research focuses on the phenomenology of electromagnetic waves as in singular optics, non-imaging optics, non-linear optics, statistical optics, and radiometry. Additionally, computer engineers have taken an interest in integrated optics, machine vision, and photonic computing as possible components of the "next generation" of computers.

Today, the pure science of optics is called optical science or optical physics to distinguish it from applied optical sciences, which are referred to as optical engineering. Prominent subfields of optical engineering include illumination engineering, photonics, and optoelectronics with practical applications like lens design, fabrication and testing of optical components, and image processing. Some of these fields overlap, with nebulous boundaries between the subjects terms that mean slightly different things in different parts of the world and in different areas of industry. A professional community of researchers in nonlinear optics has developed in the last several decades due to advances in laser technology.

Lasers



Experiments such as this one with high-power lasers are part of the modern optics research.

A laser is a device that emits light (electromagnetic radiation) through a process called *stimulated emission*. The term *laser* is an acronym for *Light Amplification by Stimulated Emission of Radiation*. Laser light is usually spatially coherent, which means that the light either is emitted in a narrow, low-divergence beam, or can be converted into one with the help of optical components such as lenses. Because the microwave equivalent of the laser, the *maser*, was developed first, devices that emit microwave and radio frequencies are usually called *masers*.

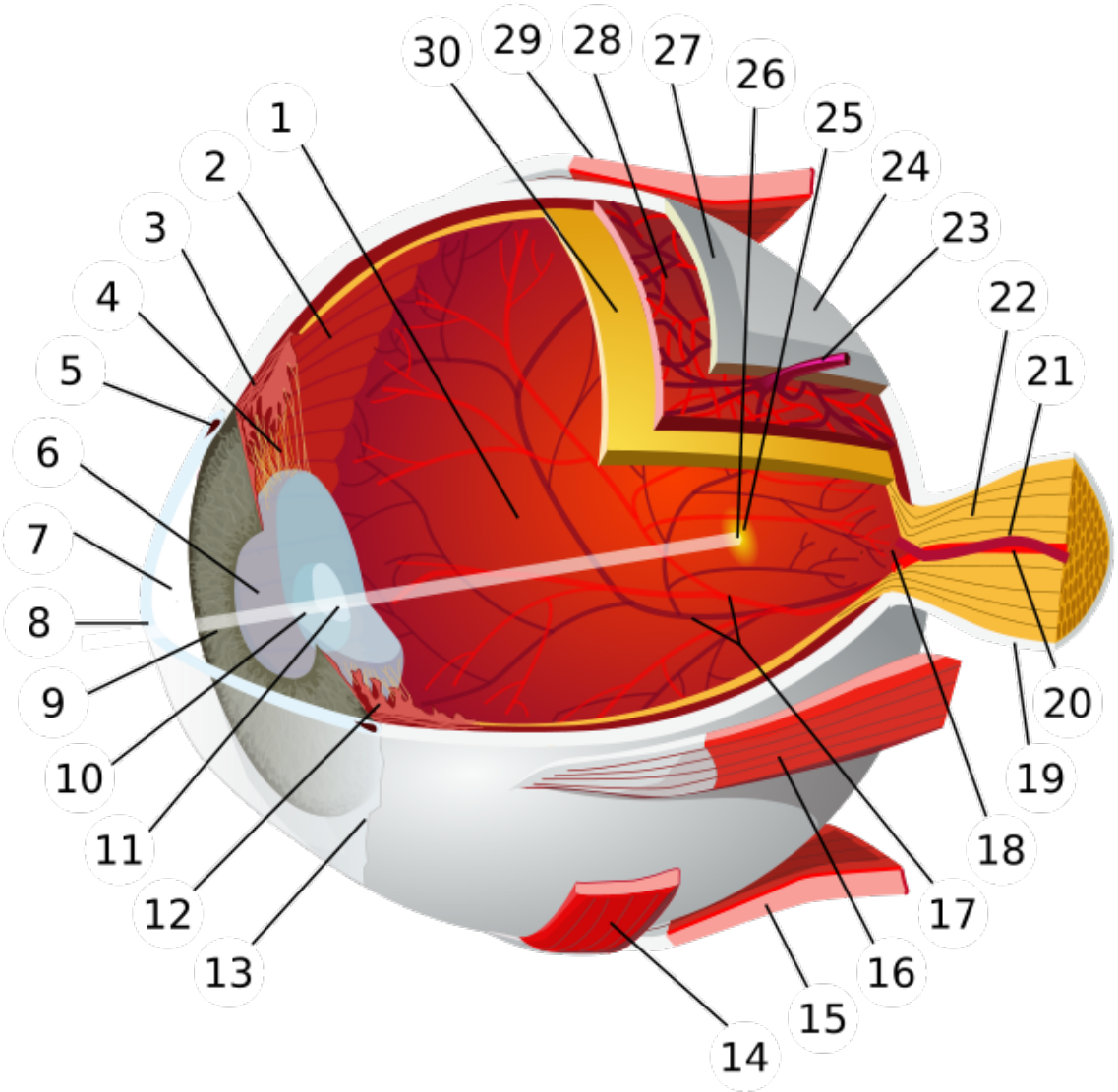
The first working laser was demonstrated on 16 May 1960 by Theodore Maiman at Hughes Research Laboratories. When first invented, they were called "a solution looking for a problem". Since then, lasers have become a multi-billion dollar industry, finding utility in thousands of highly varied applications. The first application of lasers visible in the daily lives of the general population was the supermarket barcode scanner, introduced in 1974. The laserdisc player, introduced in 1978, was the first successful consumer product to include a laser, but the compact disc player was the first laser-equipped device to become truly common in consumers' homes, beginning in 1982. These optical storage devices use a semiconductor laser less than a millimeter wide to scan the surface of the disc for data retrieval. Fiber-optic communication relies on lasers to transmit large amounts of information at the speed of light. Other common applications of lasers include laser printers and laser pointers. Lasers are used in medicine in areas such as bloodless surgery, laser eye surgery, and laser capture microdissection and in military applications

such as missile defense systems, electro-optical countermeasures (EOCM), and LIDAR. Lasers are also used in holograms, bubblegrams, laser light shows, and laser hair removal.

Applications

Optics is part of everyday life. The ubiquity of visual systems in biology indicate the central role optics plays as the science of one of the five senses. Many people benefit from eyeglasses or contact lenses, and optics are integral to the functioning of many consumer goods including cameras. Rainbows and mirages are examples of optical phenomena. Optical communication provides the backbone for both the Internet and modern telephony.

Human eye



Model of a human eye. Features mentioned here are 3. ciliary muscle, 6. pupil, 8. cornea, 10. lens cortex, 22. optic nerve, 26. fovea, 30. retina

The human eye functions by focusing light onto an array of photoreceptor cells called the retina, which covers the back of the eye. The focusing is accomplished by a series of transparent media. Light entering the eye passes first through the cornea, which provides much of the eye's optical power. The light then continues through the fluid just behind the cornea—the anterior chamber, then passes through the pupil. The light then passes through the lens, which focuses the light further and allows adjustment of focus. The light then passes through the main body of fluid in the eye—the vitreous humor, and reaches the retina. The cells in the retina cover the back of the eye, except for where the optic nerve exits; this results in a blind spot.

There are two types of photoreceptor cells, rods and cones, which are sensitive to different aspects of light. Rod cells are sensitive to the intensity of light over a wide frequency range, thus are responsible for black-and-white vision. Rod cells are not present on the fovea, the area of the retina responsible for central vision, and are not as responsive as cone cells to spatial and temporal changes in light. There are, however, twenty times more rod cells than cone cells in the retina because the rod cells are present across a wider area. Because of their wider distribution, rods are responsible for peripheral vision.

In contrast, cone cells are less sensitive to the overall intensity of light, but come in three varieties that are sensitive to different frequency-ranges and thus are used in the perception of color and photopic vision. Cone cells are highly concentrated in the fovea and have a high visual acuity meaning that they are better at spatial resolution than rod cells. Since cone cells are not as sensitive to dim light as rod cells, most night vision is limited to rod cells. Likewise, since cone cells are in the fovea, central vision (including the vision needed to do most reading, fine detail work such as sewing, or careful examination of objects) is done by cone cells.

Ciliary muscles around the lens allow the eye's focus to be adjusted. This process is known as accommodation. The near point and far point define the nearest and farthest distances from the eye at which an object can be brought into sharp focus. For a person with normal vision, the far point is located at infinity. The near point's location depends on how much the muscles can increase the curvature of the lens, and how inflexible the lens has become with age. Optometrists, ophthalmologists, and opticians usually consider an appropriate near point to be closer than normal reading distance—approximately 25 cm.

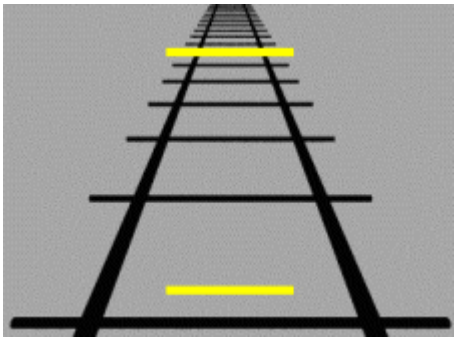
Defects in vision can be explained using optical principles. As people age, the lens becomes less flexible and the near point recedes from the eye, a condition known as presbyopia. Similarly, people suffering from hyperopia cannot decrease the focal length of their lens enough to allow for nearby objects to be imaged on their retina. Conversely, people who cannot increase the focal length of their lens enough to allow for distant objects to be imaged on the retina suffer from myopia and have a far point that is considerably closer than infinity. A condition known as astigmatism results when the cornea is not spherical but instead is more curved in one direction. This causes

horizontally extended objects to be focused on different parts of the retina than vertically extended objects, and results in distorted images.

All of these conditions can be corrected using corrective lenses. For presbyopia and hyperopia, a converging lens provides the extra curvature necessary to bring the near point closer to the eye while for myopia a diverging lens provides the curvature necessary to send the far point to infinity. Astigmatism is corrected with a cylindrical surface lens that curves more strongly in one direction than in another, compensating for the non-uniformity of the cornea.

The optical power of corrective lenses is measured in diopters, a value equal to the reciprocal of the focal length measured in meters; with a positive focal length corresponding to a converging lens and a negative focal length corresponding to a diverging lens. For lenses that correct for astigmatism as well, three numbers are given: one for the spherical power, one for the cylindrical power, and one for the angle of orientation of the astigmatism.

Visual effects



The Ponzo Illusion relies on the fact that parallel lines appear to converge as they approach infinity.

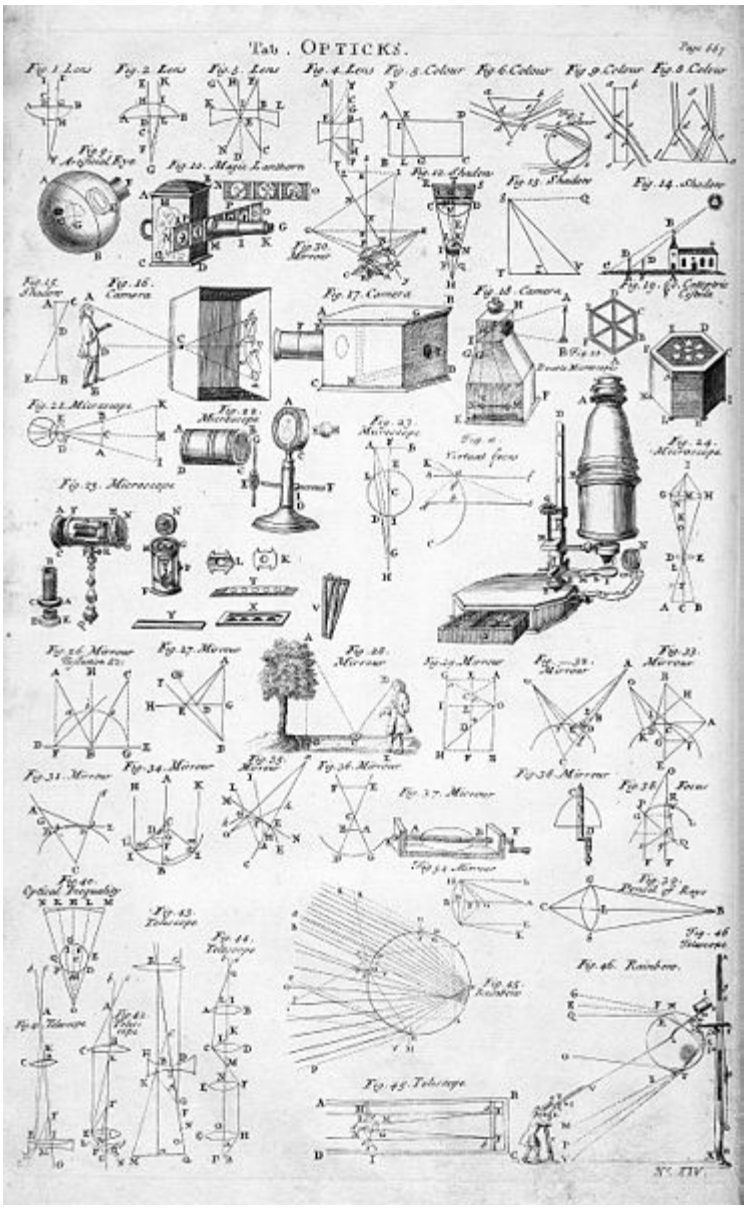
Optical illusions (also called visual illusions) are characterized by visually perceived images that differ from objective reality. The information gathered by the eye is processed in the brain to give a percept that differs from the object being imaged. Optical illusions can be the result of a variety of phenomena including physical effects that create images that are different from the objects that make them, the physiological effects on the eyes and brain of excessive stimulation (e.g. brightness, tilt, color, movement), and cognitive illusions where the eye and brain make unconscious inferences.

Cognitive illusions include some which result from the unconscious misapplication of certain optical principles. For example, the Ames room, Hering, Müller-Lyer, Orbison, Ponzo, Sander, and Wundt illusions all rely on the suggestion of the appearance of distance by using converging and diverging lines, in the same way that parallel light rays (or indeed any set of parallel lines) appear to converge at a vanishing point at infinity in two-dimensionally rendered images with artistic perspective. This suggestion is also

responsible for the famous moon illusion where the moon, despite having essentially the same angular size, appears much larger near the horizon than it does at zenith. This illusion so confounded Ptolemy that he incorrectly attributed it to atmospheric refraction when he described it in his treatise, *Optics*.

Another type of optical illusion exploits broken patterns to trick the mind into perceiving symmetries or asymmetries that are not present. Examples include the café wall, Ehrenstein, Fraser spiral, Poggendorff, and Zöllner illusions. Related, but not strictly illusions, are patterns that occur due to the superimposition of periodic structures. For example transparent tissues with a grid structure produce shapes known as moiré patterns, while the superimposition of periodic transparent patterns comprising parallel opaque lines or curves produces line moiré patterns.

Optical instruments



Illustrations of various optical instruments from the 1728 *Cyclopaedia*

Single lenses have a variety of applications including photographic lenses, corrective lenses, and magnifying glasses while single mirrors are used in parabolic reflectors and rear-view mirrors. Combining a number of mirrors, prisms, and lenses produces compound optical instruments which have practical uses. For example, a periscope is simply two plane mirrors aligned to allow for viewing around obstructions. The most famous compound optical instruments in science are the microscope and the telescope which were both invented by the Dutch in the late 16th century.

Microscopes were first developed with just two lenses: an objective lens and an eyepiece. The objective lens is essentially a magnifying glass and was designed with a very small focal length while the eyepiece generally has a longer focal length. This has the effect of producing magnified images of close objects. Generally, an additional source of illumination is used since magnified images are dimmer due to the conservation of energy and the spreading of light rays over a larger surface area. Modern microscopes, known as *compound microscopes* have many lenses in them (typically four) to optimize the functionality and enhance image stability. A slightly different variety of microscope, the comparison microscope, looks at side-by-side images to produce a stereoscopic binocular view that appears three dimensional when used by humans.

The first telescopes, called *refracting telescopes* were also developed with a single objective and eyepiece lens. In contrast to the microscope, the objective lens of the telescope was designed with a large focal length to avoid optical aberrations. The objective focuses an image of a distant object at its focal point which is adjusted to be at the focal point of an eyepiece of a much smaller focal length. The main goal of a telescope is not necessarily magnification, but rather collection of light which is determined by the physical size of the objective lens. Thus, telescopes are normally indicated by the diameters of their objectives rather than by the magnification which can be changed by switching eyepieces. Because the magnification of a telescope is equal to the focal length of the objective divided by the focal length of the eyepiece, smaller focal-length eyepieces cause greater magnification.

Since crafting large lenses is much more difficult than crafting large mirrors, most modern telescopes are *reflecting telescopes*, that is, telescopes that use a primary mirror rather than an objective lens. The same general optical considerations apply to reflecting telescopes that applied to refracting telescopes, namely, the larger the primary mirror, the more light collected, and the magnification is still equal to the focal length of the primary mirror divided by the focal length of the eyepiece. Professional telescopes generally do not have eyepieces and instead place an instrument (often a charge-coupled device) at the focal point instead.

Photography

Photograph taken with aperture *f*/32

Photograph taken with aperture *f*/5

The optics of photography involves both lenses and the medium in which the electromagnetic radiation is recorded, whether it be a plate, film, or charge-coupled device. Photographers must consider the reciprocity of the camera and the shot which is summarized by the relation

$$\text{Exposure} \propto \text{ApertureArea} \times \text{ExposureTime} \times \text{SceneLuminance}$$

In other words, the smaller the aperture (giving greater depth of focus), the less light coming in, so the length of time has to be increased (leading to possible blurriness if motion occurs). An example of the use of the law of reciprocity is the Sunny 16 rule which gives a rough estimate for the settings needed to estimate the proper exposure in daylight.

A camera's aperture is measured by a unitless number called the f-number or f-stop, ***f***/#, often notated as *N*, and given by

$$f/\# = N = \frac{f}{D}$$

where *f* is the focal length, and *D* is the diameter of the entrance pupil. By convention, "***f***/#" is treated as a single symbol, and specific values of ***f***/# are written by replacing the number sign with the value. The two ways to increase the f-stop are to either decrease the diameter of the entrance pupil or change to a longer focal length (in the case of a zoom lens, this can be done by simply adjusting the lens). Higher f-numbers also have a larger depth of field due to the lens approaching the limit of a pinhole camera which is able to focus all images perfectly, regardless of distance, but requires very long exposure times.

The field of view that the lens will provide changes with the focal length of the lens. There are three basic classifications based on the relationship to the diagonal size of the film or sensor size of the camera to the focal length of the lens:

- Normal lens: angle of view of about 50° (called *normal* because this angle considered roughly equivalent to human vision) and a focal length approximately equal to the diagonal of the film or sensor.
- Wide-angle lens: angle of view wider than 60° and focal length shorter than a normal lens.
- Long focus lens: angle of view narrow than a normal lens. This is any lens with a focal length longer than the diagonal measure of the film or sensor. The most common type of long focus lens is the telephoto lens, a design that uses a special *telephoto group* to be physically shorter than its focal length.

Modern zoom lenses may have some or all of these attributes.

The absolute value for the exposure time required depends on how sensitive to light the medium being used is (measured by the film speed, or, for digital media, by the quantum efficiency). Early photography used media that had very low light sensitivity, and so

exposure times had to be long even for very bright shots. As technology has improved, so has the sensitivity through film cameras and digital cameras.

Other results from physical and geometrical optics apply to camera optics. For example, the maximum resolution capability of a particular camera set-up is determined by the diffraction limit associated with the pupil size and given, roughly, by the Rayleigh criterion.

Atmospheric optics

A colorful sky is often due to scattering of light off particulates and pollution, as in this photograph of a sunset during the October 2007 California wildfires.

The unique optical properties of the atmosphere cause a wide range of spectacular optical phenomena. The blue color of the sky is a direct result of Rayleigh scattering which redirects higher frequency (blue) sunlight back into the field of view of the observer. Because blue light is scattered more easily than red light, the sun takes on a reddish hue when it is observed through a thick atmosphere, as during a sunrise or sunset. Additional particulate matter in the sky can scatter different colors at different angles creating colorful glowing skies at dusk and dawn. Scattering off of ice crystals and other particles in the atmosphere are responsible for halos, afterglows, coronas, rays of sunlight, and sun dogs. The variation in these kinds of phenomena is due to different particle sizes and geometries.

Mirages are optical phenomena in which light rays are bent due to thermal variations in the refraction index of air, producing displaced or heavily distorted images of distant objects. Other dramatic optical phenomena associated with this include the Novaya Zemlya effect where the sun appears to rise earlier than predicted with a distorted shape. A spectacular form of refraction occurs with a temperature inversion called the Fata Morgana where objects on the horizon or even beyond the horizon, such as islands, cliffs, ships or icebergs, appear elongated and elevated, like "fairy tale castles".

Rainbows are the result of a combination of internal reflection and dispersive refraction of light in raindrops. A single reflection off the backs of an array of raindrops produces a rainbow with an angular size on the sky that ranges from 40° to 42° with red on the outside. Double rainbows are produced by two internal reflections with angular size of 50.5° to 54° with violet on the outside. Because rainbows are seen with the sun 180° away from the center of the rainbow, rainbows are more prominent the closer the sun is to the horizon.

Chapter 9

Energy

Lightning is the electric breakdown of air by strong electric fields, which produce a force on charges. When these charges move through a distance, a flow of energy occurs. The electric potential energy in the atmosphere then is transformed into thermal energy, light, and sound, which are other forms of energy.

In physics, **energy** is a quantity that is often understood as the ability a physical system has to do work on other physical systems. Since work is defined as a force acting through a distance (a length of space), energy is always equivalent to the ability to exert pulls or pushes against the basic forces of nature, along a path of a certain length.

The total energy contained in an object is identified with its mass, and energy (like mass), cannot be created or destroyed. When matter (ordinary material particles) is changed into energy (such as energy of motion, or into radiation), the **mass** of the system does not change through the transformation process. However, there may be mechanistic limits as to how much of the matter in an object may be changed into other types of energy and thus into work, on other systems. Energy, like mass, is a scalar physical quantity. In the International System of Units (SI), energy is measured in joules, but in many fields other units, such as kilowatt-hours and kilocalories, are customary. All of these units translate to units of work, which is always defined in terms of forces and the distances that the forces act through.

A system can transfer energy to another system by simply transferring matter to it (since matter is equivalent to energy, in accordance with its mass). However, when energy is transferred by means other than matter-transfer, the transfer produces changes in the second system, as a result of work done on it. This work manifests itself as the effect of force(s) applied through distances within the target system. For example, a system can emit energy to another by transferring (radiating) electromagnetic energy, but this creates forces upon the particles that absorb the radiation. Similarly, a system may transfer energy to another by physically impacting it, but that case the energy of motion in an object, called kinetic energy, results in forces acting over distances (new energy) to appear in another object that is struck. Transfer of thermal energy by heat occurs by both of these mechanisms: heat can be transferred by electromagnetic radiation, or by physical contact in which direct particle-particle impacts transfer kinetic energy.

Energy may be stored in systems without being present as matter, or as kinetic or electromagnetic energy. Stored energy is created whenever a particle has been moved through a field it interacts with (requiring a force to do so), but the energy to accomplish this is stored as a new position of the particles in the field—a configuration that must be "held" or fixed by a different type of force (otherwise, the new configuration would resolve itself by the field pushing or pulling the particle back toward its previous position). This type of energy "stored" by force-fields and particles that have been forced into a new physical configuration in the field by doing work on them by another system, is referred to as potential energy. A simple example of potential energy is the work needed to lift an object in a gravity field, up to a support. Each of the basic forces of nature is associated with a different type of potential energy, and all types of potential energy (like all other types of energy) appears as system mass, whenever present. For example, a compressed spring will be slightly more massive than before it was compressed. Likewise, whenever energy is transferred between systems by any mechanism, an associated mass is transferred with it.

Any form of energy may be transformed into another form. For example, all types of potential energy are converted into kinetic energy when the objects are given freedom to move to different position (as for example, when an object falls off a support). When energy is in a form other than thermal energy, it may be transformed with good or even perfect efficiency, to any other type of energy, including electricity or production of new particles of matter. With thermal energy, however, there are often limits to the efficiency

of the conversion to other forms of energy, as described by the second law of thermodynamics.

In all such energy transformation processes, the total energy remains the same, and a transfer of energy from one system to another, results in a loss to compensate for any gain. This principle, the conservation of energy, was first postulated in the early 19th century, and applies to any isolated system. According to Noether's theorem, the conservation of energy is a consequence of the fact that the laws of physics do not change over time.

Although the total energy of a system does not change with time, its value may depend on the frame of reference. For example, a seated passenger in a moving airplane has zero kinetic energy relative to the airplane, but non-zero kinetic energy (and higher total energy) relative to the Earth.

History

The word *energy* derives from the Greek ἐνέργεια *energeia*, which possibly appears for the first time in the work of Aristotle in the 4th century BC.

The concept of energy emerged out of the idea of *vis viva* (living force), which Gottfried Leibniz defined as the product of the mass of an object and its velocity squared; he believed that total *vis viva* was conserved. To account for slowing due to friction, Leibniz theorized that thermal energy consisted of the random motion of the constituent parts of matter, a view shared by Isaac Newton, although it would be more than a century until this was generally accepted. In 1807, Thomas Young was possibly the first to use the term "energy" instead of *vis viva*, in its modern sense. Gustave-Gaspard Coriolis described "kinetic energy" in 1829 in its modern sense, and in 1853, William Rankine coined the term "potential energy". It was argued for some years whether energy was a substance (the caloric) or merely a physical quantity, such as momentum.

William Thomson (Lord Kelvin) amalgamated all of these laws into the laws of thermodynamics, which aided in the rapid development of explanations of chemical processes by Rudolf Clausius, Josiah Willard Gibbs, and Walther Nernst. It also led to a mathematical formulation of the concept of entropy by Clausius and to the introduction of laws of radiant energy by Jožef Stefan.

During a 1961 lecture for undergraduate students at the California Institute of Technology, Richard Feynman, a celebrated physics teacher and Nobel Laureate, said this about the concept of energy:

There is a fact, or if you wish, a *law*, governing all natural phenomena that are known to date. There is no known exception to this law—it is exact so far as we know. The law is called the *conservation of energy*. It states that there is a certain quantity, which we call energy, that does not change in manifold changes which nature undergoes. That is a most abstract idea, because it is a mathematical principle; it says that there is a numerical

quantity which does not change when something happens. It is not a description of a mechanism, or anything concrete; it is just a strange fact that we can calculate some number and when we finish watching nature go through her tricks and calculate the number again, it is the same.

Since 1918 it has been known that the law of conservation of energy is the direct mathematical consequence of the translational symmetry of the quantity conjugate to energy, namely time. That is, energy is conserved because the laws of physics do not distinguish between different instants of time.

Energy in various contexts

The concept of energy and its transformations is useful in explaining and predicting most natural phenomena. The *direction* of transformations in energy (what kind of energy is transformed to what other kind) is often described by entropy (equal energy spread among all available degrees of freedom) considerations, as in practice all energy transformations are permitted on a small scale, but certain larger transformations are not permitted because it is statistically unlikely that energy or matter will randomly move into more concentrated forms or smaller spaces.

The concept of energy is widespread in all sciences.

- In the context of chemistry, energy is an attribute of a substance as a consequence of its atomic, molecular or aggregate structure. Since a chemical transformation is accompanied by a change in one or more of these kinds of structure, it is invariably accompanied by an increase or decrease of energy of the substances involved. Some energy is transferred between the surroundings and the reactants of the reaction in the form of heat or light; thus the products of a reaction may have more or less energy than the reactants. A reaction is said to be exergonic if the final state is lower on the energy scale than the initial state; in the case of endergonic reactions the situation is the reverse. Chemical reactions are invariably not possible unless the reactants surmount an energy barrier known as the activation energy. The *speed* of a chemical reaction (at given temperature T) is related to the activation energy E, by the Boltzmann's population factor $e^{-E/kT}$ – that is the probability of molecule to have energy greater than or equal to E at the given temperature T. This exponential dependence of a reaction rate on temperature is known as the Arrhenius equation. The activation energy necessary for a chemical reaction can be in the form of thermal energy.
- In biology, energy is an attribute of all biological systems from the biosphere to the smallest living organism. Within an organism it is responsible for growth and development of a biological cell or an organelle of a biological organism. Energy is thus often said to be stored by cells in the structures of molecules of substances such as carbohydrates (including sugars), lipids, and proteins, which release energy when reacted with oxygen in respiration. In human terms, the human equivalent (H-e) (Human energy conversion) indicates, for a given amount of energy expenditure, the relative quantity of energy needed for human metabolism,

assuming an average human energy expenditure of 12,500kJ per day and a basal metabolic rate of 80 watts. For example, if our bodies run (on average) at 80 watts, then a light bulb running at 100 watts is running at 1.25 human equivalents (100 ÷ 80) i.e. 1.25 H-e. For a difficult task of only a few seconds' duration, a person can put out thousands of watts, many times the 746 watts in one official horsepower. For tasks lasting a few minutes, a fit human can generate perhaps 1,000 watts. For an activity that must be sustained for an hour, output drops to around 300; for an activity kept up all day, 150 watts is about the maximum. The human equivalent assists understanding of energy flows in physical and biological systems by expressing energy units in human terms: it provides a "feel" for the use of a given amount of energy

- In geology, continental drift, mountain ranges, volcanoes, and earthquakes are phenomena that can be explained in terms of energy transformations in the Earth's interior., while meteorological phenomena like wind, rain, hail, snow, lightning, tornadoes and hurricanes, are all a result of energy transformations brought about by solar energy on the atmosphere of the planet Earth.
- In cosmology and astronomy the phenomena of stars, nova, supernova, quasars and gamma ray bursts are the universe's highest-output energy transformations of matter. All stellar phenomena (including solar activity) are driven by various kinds of energy transformations. Energy in such transformations is either from gravitational collapse of matter (usually molecular hydrogen) into various classes of astronomical objects (stars, black holes, etc.), or from nuclear fusion (of lighter elements, primarily hydrogen).

Energy transformations in the universe over time are characterized by various kinds of potential energy that has been available since the Big Bang, later being "released" (transformed to more active types of energy such as kinetic or radiant energy), when a triggering mechanism is available.

Familiar examples of such processes include nuclear decay, in which energy is released that was originally "stored" in heavy isotopes (such as uranium and thorium), by nucleosynthesis, a process ultimately using the gravitational potential energy released from the gravitational collapse of supernovae, to store energy in the creation of these heavy elements before they were incorporated into the solar system and the Earth. This energy is triggered and released in nuclear fission bombs. In a slower process, **radioactive decay** of these atoms in the core of the Earth releases heat. This thermal energy drives plate tectonics and may lift mountains, via orogenesis. This slow lifting represents a kind of gravitational potential energy storage of the thermal energy, which may be later released to active kinetic energy in landslides, after a triggering event. Earthquakes also release stored elastic potential energy in rocks, a store that has been produced ultimately from the same radioactive heat sources. Thus, according to present understanding, familiar events such as landslides and earthquakes release energy that has been stored as potential energy in the Earth's gravitational field or elastic strain (mechanical potential energy) in rocks. Prior to this, they represent release of energy that has been stored in heavy atoms since the collapse of long-destroyed supernova stars created these atoms.

In another similar chain of transformations beginning at the dawn of the universe, **nuclear fusion** of hydrogen in the Sun also releases another store of potential energy which was created at the time of the Big Bang. At that time, according to theory, space expanded and the universe cooled too rapidly for hydrogen to completely fuse into heavier elements. This meant that hydrogen represents a store of potential energy that can be released by fusion. Such a fusion process is triggered by heat and pressure generated from gravitational collapse of hydrogen clouds when they produce stars, and some of the fusion energy is then transformed into sunlight. Such sunlight from our Sun may again be stored as gravitational potential energy after it strikes the Earth, as (for example) water evaporates from oceans and is deposited upon mountains (where, after being released at a hydroelectric dam, it can be used to drive turbines or generators to produce electricity). Sunlight also drives many weather phenomena, save those generated by volcanic events. An example of a solar-mediated weather event is a hurricane, which occurs when large unstable areas of warm ocean, heated over months, give up some of their thermal energy suddenly to power a few days of violent air movement. Sunlight is also captured by plants as *chemical potential energy* in photosynthesis, when carbon dioxide and water (two low-energy compounds) are converted into the high-energy compounds carbohydrates, lipids, and proteins. Plants also release oxygen during photosynthesis, which is utilized by living organisms as an electron acceptor, to release the energy of carbohydrates, lipids, and proteins. Release of the energy stored during photosynthesis as heat or light may be triggered suddenly by a spark, in a forest fire, or it may be made available more slowly for animal or human metabolism, when these molecules are ingested, and catabolism is triggered by enzyme action.

Through all of these transformation chains, potential energy stored at the time of the Big Bang is later released by intermediate events, sometimes being stored in a number of ways over time between releases, as more active energy. In all these events, one kind of energy is converted to other types of energy, including heat.

Conservation of energy

Energy is subject to the **law of conservation of energy**. According to this law, energy can neither be created (produced) nor destroyed by itself. It can only be transformed.

Most kinds of energy (with gravitational energy being a notable exception) are also subject to strict local conservation laws, as well. In this case, energy can only be exchanged between adjacent regions of space, and all observers agree as to the volumetric density of energy in any given space. There is also a global law of conservation of energy, stating that the total energy of the universe cannot change; this is a corollary of the local law, but not vice versa. Conservation of energy is the mathematical consequence of translational symmetry of time (that is, the indistinguishability of time intervals taken at different time).

According to energy conservation law the total inflow of energy into a system must equal the total outflow of energy from the system, plus the change in the energy contained within the system.

This law is a fundamental principle of physics. It follows from the translational symmetry of time, a property of most phenomena below the cosmic scale that makes them independent of their locations on the time coordinate. Put differently, yesterday, today, and tomorrow are physically indistinguishable.

This is because energy is the quantity which is canonical conjugate to time. This mathematical entanglement of energy and time also results in the uncertainty principle - it is impossible to define the exact amount of energy during any definite time interval. The uncertainty principle should not be confused with energy conservation - rather it provides mathematical limits to which energy can in principle be defined and measured.

In quantum mechanics energy is expressed using the Hamiltonian operator. On any time scales, the uncertainty in the energy is by

$$\Delta E \Delta t \geq \frac{\hbar}{2}$$

which is similar in form to the Heisenberg uncertainty principle (but not really mathematically equivalent thereto, since H and t are not dynamically conjugate variables, neither in classical nor in quantum mechanics).

In particle physics, this inequality permits a qualitative understanding of virtual particles which carry momentum, exchange by which and with real particles, is responsible for the creation of all known fundamental forces (more accurately known as fundamental interactions). Virtual photons (which are simply lowest quantum mechanical energy state of photons) are also responsible for electrostatic interaction between electric charges (which results in Coulomb law), for spontaneous radiative decay of exited atomic and nuclear states, for the Casimir force, for van der Waals bond forces and some other observable phenomena.

Applications of the concept of energy

Energy is subject to a strict global conservation law; that is, whenever one measures (or calculates) the total energy of a system of particles whose interactions do not depend explicitly on time, it is found that the total energy of the system always remains constant.

- The total energy of a system can be subdivided and classified in various ways. For example, it is sometimes convenient to distinguish potential energy (which is a function of coordinates only) from kinetic energy (which is a function of coordinate time derivatives only). It may also be convenient to distinguish gravitational energy, electric energy, thermal energy, and other forms. These classifications overlap; for instance, thermal energy usually consists partly of kinetic and partly of potential energy.
- The *transfer* of energy can take various forms; familiar examples include work, heat flow, and advection, as discussed below.

- The word "energy" is also used outside of physics in many ways, which can lead to ambiguity and inconsistency. The vernacular terminology is not consistent with technical terminology. For example, while energy is always conserved (in the sense that the total energy does not change despite energy transformations), energy can be converted into a form, e.g., thermal energy, that cannot be utilized to perform work. When one talks about "conserving energy by driving less," one talks about conserving fossil fuels and preventing useful energy from being lost as heat. This usage of "conserve" differs from that of the law of conservation of energy.

In classical physics energy is considered a scalar quantity, the canonical conjugate to time. In special relativity energy is also a scalar (although not a Lorentz scalar but a time component of the energy-momentum 4-vector). In other words, energy is invariant with respect to rotations of space, but not invariant with respect to rotations of space-time (= boosts).

Energy transfer

Because energy is strictly conserved and is also locally conserved (wherever it can be defined), it is important to remember that by the definition of energy the transfer of energy between the "system" and adjacent regions is work. A familiar example is *mechanical work*. In simple cases this is written as the following equation:

$$\Delta E = W \qquad (1)$$

if there are no other energy-transfer processes involved. Here E is the amount of energy transferred, and W represents the work done on the system.

More generally, the energy transfer can be split into two categories:

$$\Delta E = W + Q \qquad (2)$$

where Q represents the heat flow into the system.

There are other ways in which an open system can gain or lose energy. In chemical systems, energy can be added to a system by means of adding substances with different chemical potentials, which potentials are then extracted (both of these process are illustrated by fueling an auto, a system which gains in energy thereby, without addition of either work or heat). Winding a clock would be adding energy to a mechanical system. These terms may be added to the above equation, or they can generally be subsumed into a quantity called "energy addition term E" which refers to *any* type of energy carried over the surface of a control volume or system volume. Examples may be seen above, and many others can be imagined (for example, the kinetic energy of a stream of particles entering a system, or energy from a laser beam adds to system energy, without either being either work-done or heat-added, in the classic senses).

$$\Delta E = W + Q + E \quad (3)$$

Where E in this general equation represents other additional advected energy terms not covered by work done on a system, or heat added to it.

Energy is also transferred from potential energy (E_p) to kinetic energy (E_k) and then back to potential energy constantly. This is referred to as conservation of energy. In this closed system, energy cannot be created or destroyed; therefore, the initial energy and the final energy will be equal to each other. This can be demonstrated by the following:

$$E_{pi} + E_{ki} = E_{pF} + E_{kF} \quad (4)$$

The equation can then be simplified further since $E_p = mgh$ (mass times acceleration due to gravity times the height) and $E_k = \frac{1}{2}mv^2$ (half mass times velocity squared). Then the total amount of energy can be found by adding $E_p + E_k = E_{total}$.

Energy and the laws of motion

In classical mechanics, energy is a conceptually and mathematically useful property, as it is a conserved quantity. Several formulations of mechanics have been developed using energy as a core concept.

The Hamiltonian

The total energy of a system is sometimes called the Hamiltonian, after William Rowan Hamilton. The classical equations of motion can be written in terms of the Hamiltonian, even for highly complex or abstract systems. These classical equations have remarkably direct analogs in nonrelativistic quantum mechanics.

The Lagrangian

Another energy-related concept is called the Lagrangian, after Joseph Louis Lagrange. This is even more fundamental than the Hamiltonian, and can be used to derive the equations of motion. It was invented in the context of classical mechanics, but is generally useful in modern physics. The Lagrangian is defined as the kinetic energy *minus* the potential energy.

Usually, the Lagrange formalism is mathematically more convenient than the Hamiltonian for non-conservative systems (such as systems with friction).

Energy and thermodynamics

Internal energy

Internal energy is the sum of all microscopic forms of energy of a system. It is the energy needed to create the system. It is related to the potential energy, e.g., molecular structure, crystal structure, and other geometric aspects, as well as the motion of the particles, in form of kinetic energy. Thermodynamics is chiefly concerned with changes in internal energy and not its absolute value, which is impossible to determine with thermodynamics alone.

The laws of thermodynamics

According to the second law of thermodynamics, work can be totally converted into heat, but not vice versa. This is a mathematical consequence of statistical mechanics. The first law of thermodynamics simply asserts that energy is conserved, and that heat is included as a form of energy transfer. A commonly used corollary of the first law is that for a "system" subject only to pressure forces and heat transfer (e.g., a cylinder-full of gas), the differential change in energy of the system (with a *gain* in energy signified by a positive quantity) is given as the following equation:

$$\mathrm{d}E = T\mathrm{d}S - P\mathrm{d}V,$$

where the first term on the right is the heat transfer into the system, defined in terms of temperature T and entropy S (in which entropy increases and the change dS is positive when the system is heated), and the last term on the right hand side is identified as "work" done on the system, where pressure is P and volume V (the negative sign results since compression of the system requires work to be done on it and so the volume change, dV, is negative when work is done on the system). Although this equation is the standard textbook example of energy conservation in classical thermodynamics, it is highly specific, ignoring all chemical, electric, nuclear, and gravitational forces, effects such as advection of any form of energy other than heat, and because it contains a term that depends on temperature. The most general statement of the first law (i.e., conservation of energy) is valid even in situations in which temperature is undefinable.

Energy is sometimes expressed as the following equation:

$$\mathrm{d}E = \delta Q + \delta W,$$

which is unsatisfactory because there cannot exist any thermodynamic state functions W or Q that are meaningful on the right hand side of this equation, except perhaps in trivial cases.

Equipartition of energy

The energy of a mechanical harmonic oscillator (a mass on a spring) is alternatively kinetic and potential. At two points in the oscillation cycle it is entirely kinetic, and alternatively at two other points it is entirely potential. Over the whole cycle, or over many cycles, net energy is thus equally split between kinetic and potential. This is called equipartition principle; total energy of a system with many degrees of freedom is equally split among all available degrees of freedom.

This principle is vitally important to understanding the behavior of a quantity closely related to energy, called entropy. Entropy is a measure of evenness of a distribution of energy between parts of a system. When an isolated system is given more degrees of freedom (i.e., given new available energy states that are the same as existing states), then total energy spreads over **all** available degrees equally without distinction between "new" and "old" degrees. This mathematical result is called the second law of thermodynamics.

Oscillators, phonons, and photons

In an ensemble (connected collection) of unsynchronized oscillators, the average energy is spread equally between kinetic and potential types.

In a solid, thermal energy (often referred to loosely as heat content) can be accurately described by an ensemble of thermal phonons that act as mechanical oscillators. In this model, thermal energy is equally kinetic and potential.

In an ideal gas, the interaction potential between particles is essentially the delta function which stores no energy: thus, all of the thermal energy is kinetic.

Because an electric oscillator (LC circuit) is analogous to a mechanical oscillator, its energy must be, on average, equally kinetic and potential. It is entirely arbitrary whether the magnetic energy is considered kinetic and whether the electric energy is considered potential, or vice versa. That is, either the inductor is analogous to the mass while the capacitor is analogous to the spring, or vice versa.

1. By extension of the previous line of thought, in free space the electromagnetic field can be considered an ensemble of oscillators, meaning that radiation energy can be considered equally potential and kinetic. This model is useful, for example, when the electromagnetic Lagrangian is of primary interest and is interpreted in terms of potential and kinetic energy.

2. On the other hand, in the key equation $m^2c^4 = E^2 - p^2c^2$, the contribution mc^2 is called the rest energy, and all other contributions to the energy are called kinetic energy. For a particle that has mass, this implies that the kinetic energy is $0.5p^2 / m$ at speeds much smaller than c, as can be proved by writing $E = mc^2 \sqrt{(1 + p^2m^{-2}c^{-2})}$ and expanding the square root to lowest order. By this line of reasoning, the energy of a photon is entirely

kinetic, because the photon is massless and has no rest energy. This expression is useful, for example, when the energy-versus-momentum relationship is of primary interest.

The two analyses are entirely consistent. The electric and magnetic degrees of freedom in item 1 are *transverse* to the direction of motion, while the speed in item 2 is *along* the direction of motion. For non-relativistic particles these two notions of potential versus kinetic energy are numerically equal, so the ambiguity is harmless, but not so for relativistic particles.

Work and virtual work

Work, a form of energy, is force times distance.

$$W = \int_C \mathbf{F} \cdot \mathrm{d}\mathbf{s}$$

This says that the work (W) is equal to the line integral of the force $\mathbf{F}$ along a path C.

Work and thus energy is frame dependent. For example, consider a ball being hit by a bat. In the center-of-mass reference frame, the bat does no work on the ball. But, in the reference frame of the person swinging the bat, considerable work is done on the ball.

Quantum mechanics

In quantum mechanics energy is defined in terms of the energy operator as a time derivative of the wave function. The Schrödinger equation equates the energy operator to the full energy of a particle or a system. In results can be considered as a definition of measurement of energy in quantum mechanics. The Schrödinger equation describes the space- and time-dependence of slow changing (non-relativistic) wave function of quantum systems. The solution of this equation for bound system is discrete (a set of permitted states, each characterized by an energy level) which results in the concept of quanta. In the solution of the Schrödinger equation for any oscillator (vibrator) and for electromagnetic waves in a vacuum, the resulting energy states are related to the frequency by the Planck equation $E = h\nu$ (where h is the Planck's constant and ν the frequency). In the case of electromagnetic wave these energy states are called quanta of light or photons.

Relativity

When calculating kinetic energy (work to accelerate a mass from zero speed to some finite speed) relativistically - using Lorentz transformations instead of Newtonian mechanics, Einstein discovered an unexpected by-product of these calculations to be an energy term which does not vanish at zero speed. He called it rest mass energy - energy which every mass must possess even when being at rest. The amount of energy is directly proportional to the mass of body:

$$E = mc^2,$$

where

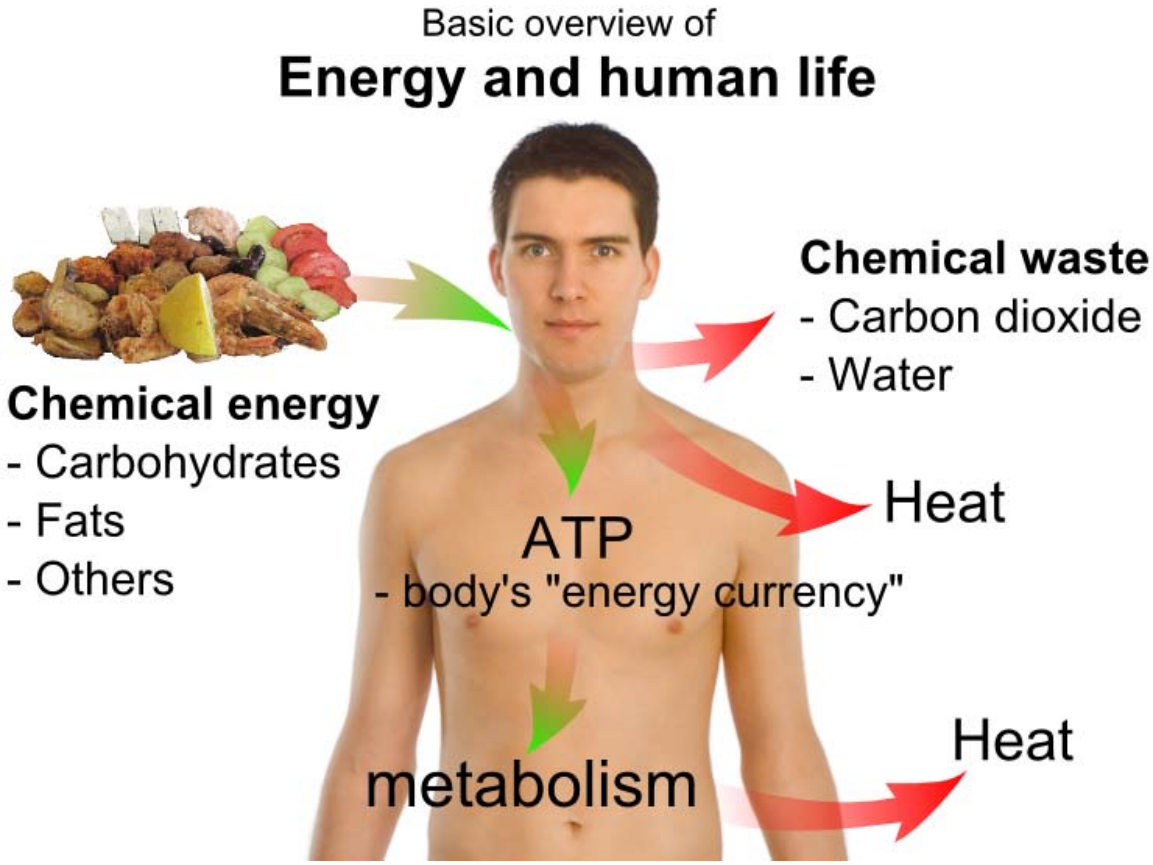
m is the mass,
c is the speed of light in vacuum,
E is the rest mass energy.

For example, consider electron-positron annihilation, in which the rest mass of individual particles is destroyed, but the inertia equivalent of the system of the two particles (its invariant mass) remains (since all energy is associated with mass), and this inertia and invariant mass is carried off by photons which individually are massless, but as a system retain their mass. This is a reversible process - the inverse process is called pair creation - in which the rest mass of particles is created from energy of two (or more) annihilating photons. In this system the matter (electrons and positrons) is destroyed and changed to non-matter energy (the photons). However, the total system mass and energy do not change during this interaction.

In general relativity, the stress-energy tensor serves as the source term for the gravitational field, in rough analogy to the way mass serves as the source term in the non-relativistic Newtonian approximation.

It is not uncommon to hear that energy is "equivalent" to mass. It would be more accurate to state that every energy has an inertia and gravity equivalent, and because mass is a form of energy, then mass too has inertia and gravity associated with it.

Energy and life



Basic overview of energy and human life.

Any living organism relies on an external source of energy—radiation from the Sun in the case of green plants; chemical energy in some form in the case of animals—to be able to grow and reproduce. The daily 1500–2000 Calories (6–8 MJ) recommended for a human adult are taken as a combination of oxygen and food molecules, the latter mostly carbohydrates and fats, of which glucose ($C_6H_{12}O_6$) and stearin ($C_{57}H_{110}O_6$) are convenient examples. The food molecules are oxidised to carbon dioxide and water in the mitochondria

$$C_6H_{12}O_6 + 6O_2 \rightarrow 6CO_2 + 6H_2O$$
$$C_{57}H_{110}O_6 + 81.5O_2 \rightarrow 57CO_2 + 55H_2O$$

and some of the energy is used to convert ADP into ATP

$$ADP + HPO_4^{2-} \rightarrow ATP + H_2O$$

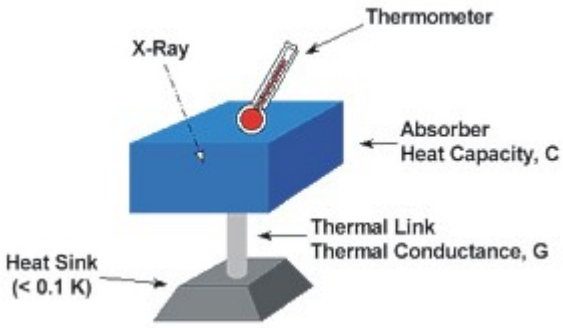
The rest of the chemical energy in the carbohydrate or fat is converted into heat: the ATP is used as a sort of "energy currency", and some of the chemical energy it contains when split and reacted with water, is used for other metabolism (at each stage of a metabolic

pathway, some chemical energy is converted into heat). Only a tiny fraction of the original chemical energy is used for work:

gain in kinetic energy of a sprinter during a 100 m race: 4 kJ
gain in gravitational potential energy of a 150 kg weight lifted through 2 metres: 3kJ
Daily food intake of a normal adult: 6–8 MJ

It would appear that living organisms are remarkably inefficient (in the physical sense) in their use of the energy they receive (chemical energy or radiation), and it is true that most real machines manage higher efficiencies. In growing organisms the energy that is converted to heat serves a vital purpose, as it allows the organism tissue to be highly ordered with regard to the molecules it is built from. The second law of thermodynamics states that energy (and matter) tends to become more evenly spread out across the universe: to concentrate energy (or matter) in one specific place, it is necessary to spread out a greater amount of energy (as heat) across the remainder of the universe ("the surroundings"). Simpler organisms can achieve higher energy efficiencies than more complex ones, but the complex organisms can occupy ecological niches that are not available to their simpler brethren. The conversion of a portion of the chemical energy to heat at each step in a metabolic pathway is the physical reason behind the pyramid of biomass observed in ecology: to take just the first step in the food chain, of the estimated 124.7 Pg/a of carbon that is fixed by photosynthesis, 64.3 Pg/a (52%) are used for the metabolism of green plants, i.e. reconverted into carbon dioxide and heat.

Measurement



A schematic diagram of a Calorimeter - An instrument used by physicists to measure energy

Because energy is defined as the ability to do work on objects, there is no absolute measure of energy. Only the transition of a system from one state into another can be defined and thus energy is measured in relative terms. The choice of a baseline or zero point is often arbitrary and can be made in whatever way is most convenient for a problem.

Methods

The methods for the measurement of energy often deploy methods for the measurement of still more fundamental concepts of science, namely mass, distance, radiation, temperature, time, electric charge and electric current.

Conventionally the technique most often employed is calorimetry, a thermodynamic technique that relies on the measurement of temperature using a thermometer or of intensity of radiation using a bolometer.

Units

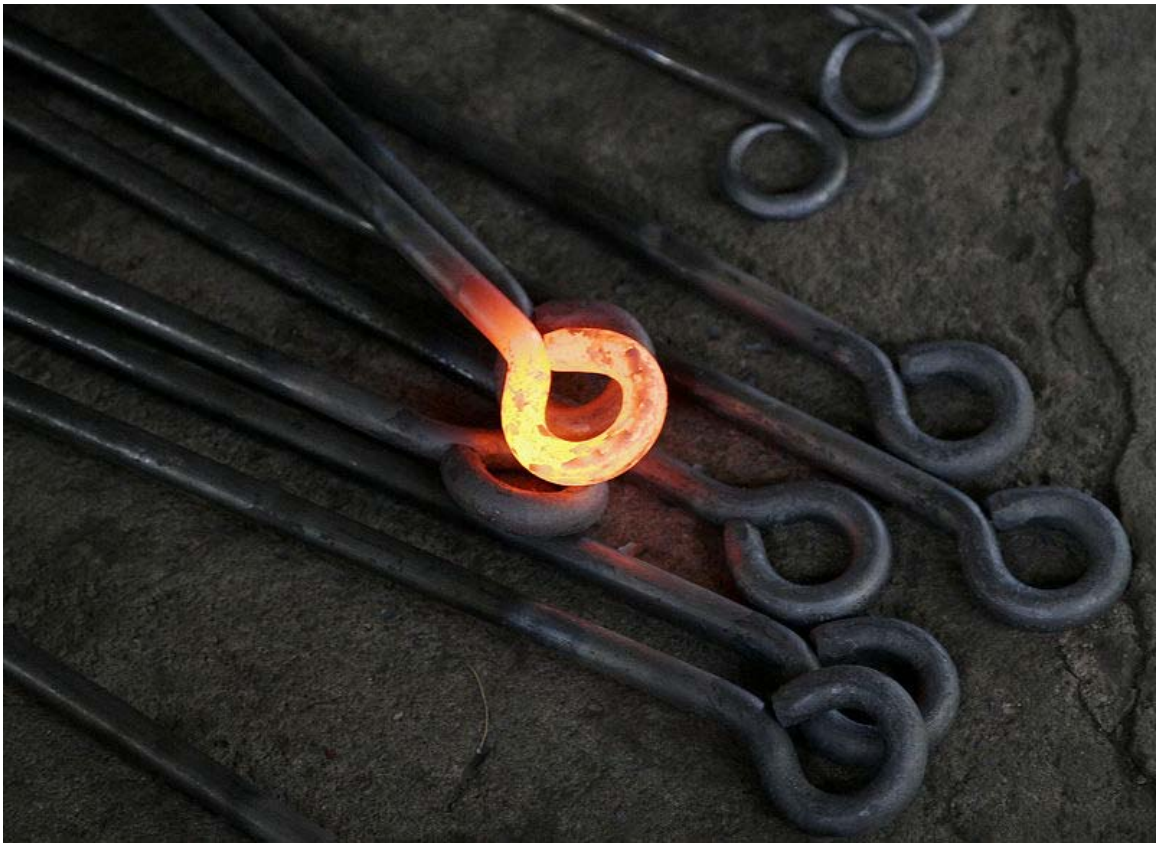
Throughout the history of science, energy has been expressed in several different units such as ergs and calories. At present, the accepted unit of measurement for energy is the SI unit of energy, the joule. In addition to the joule, other units of energy include the kilowatt hour (kWh) and the British thermal unit (Btu). These are both larger units of energy. One kWh is equivalent to exactly 3.6 million joules, and one Btu is equivalent to about 1055 joules.

Energy density

Energy density is a term used for the amount of useful energy stored in a given system or region of space per unit volume.

For fuels, the energy per unit volume is sometimes a useful parameter. In a few applications, comparing, for example, the effectiveness of hydrogen fuel to gasoline it turns out that hydrogen has a higher specific energy than does gasoline, but, even in liquid form, a much lower energy *density*.

Forms of energy



Heat, a form of energy, is partly potential energy and partly kinetic energy.

In the context of physical sciences, several forms of energy have been defined. These include:

- Thermal energy, thermal energy in transit is called heat
- Chemical energy
- Electrical energy
- Radiant energy, the energy of electromagnetic radiation
- Nuclear energy
- Magnetic energy
- Elastic energy
- Sound energy
- Mechanical energy
- Luminous energy

These energies may be divided into two main groups; kinetic energy and potential energy. Other familiar types of energy are a varying mix of both potential and kinetic energy. Energy may be transformed between these forms.

The above list of the known possible forms of energy is not necessarily complete. Whenever physical scientists discover that a certain phenomenon appears to violate the law of energy conservation, new forms may be added, as is the case with dark energy, a

hypothetical form of energy that permeates all of space and tends to increase the rate of expansion of the universe.

Classical mechanics distinguishes between potential energy, which is a function of the position of an object, and kinetic energy, which is a function of its movement. Both position and movement are relative to a frame of reference, which must be specified: this is often (and originally) an arbitrary fixed point on the surface of the Earth, the *terrestrial* frame of reference. It has been attempted to categorize *all* forms of energy as either kinetic or potential: this is not incorrect, but neither is it clear that it is a real simplification, as Feynman points out:

These notions of potential and kinetic energy depend on a notion of length scale. For example, one can speak of *macroscopic* potential and kinetic energy, which do not include thermal potential and kinetic energy. Also what is called chemical potential energy (below) is a macroscopic notion, and closer examination shows that it is really the sum of the potential *and kinetic* energy on the atomic and subatomic scale. Similar remarks apply to nuclear "potential" energy and most other forms of energy. This dependence on length scale is non-problematic if the various length scales are decoupled, as is often the case ... but confusion can arise when different length scales are coupled, for instance when friction converts macroscopic work into microscopic thermal energy.

Transformations of energy

One form of energy can often be readily transformed into another with the help of a device- for instance, a battery, from chemical energy to electric energy; a dam: gravitational potential energy to kinetic energy of moving water (and the blades of a turbine) and ultimately to electric energy through an electric generator. Similarly, in the case of a chemical explosion, chemical potential energy is transformed to kinetic energy and thermal energy in a very short time. Yet another example is that of a pendulum. At its highest points the kinetic energy is zero and the gravitational potential energy is at maximum. At its lowest point the kinetic energy is at maximum and is equal to the decrease of potential energy. If one (unrealistically) assumes that there is no friction, the conversion of energy between these processes is perfect, and the pendulum will continue swinging forever.

Energy gives rise to weight when it is trapped in a system with zero momentum, where it can be weighed. It is also equivalent to mass, and this mass is always associated with it. Mass is also equivalent to a certain amount of energy, and likewise always appears associated with it, as described in mass-energy equivalence. The formula $E = mc^2$, derived by Albert Einstein (1905) quantifies the relationship between rest-mass and rest-energy within the concept of special relativity. In different theoretical frameworks, similar formulas were derived by J. J.

Matter may be destroyed and converted to energy (and vice versa), but mass cannot ever be destroyed; rather, mass remains a constant for both the matter and the energy, during any process when they are converted into each other. However, since c^2 is extremely

large relative to ordinary human scales, the conversion of ordinary amount of matter (for example, 1 kg) to other forms of energy (such as heat, light, and other radiation) can liberate tremendous amounts of energy (~$9x10^{16}$ joules = 21 megatons of TNT), as can be seen in nuclear reactors and nuclear weapons. Conversely, the mass equivalent of a unit of energy is minuscule, which is why a loss of energy (loss of mass) from most systems is difficult to measure by weight, unless the energy loss is very large. Examples of energy transformation into matter (i.e., kinetic energy into particles with rest mass) are found in high-energy nuclear physics.

Transformation of energy into useful work is a core topic of thermodynamics. In nature, transformations of energy can be fundamentally classed into two kinds: those that are thermodynamically reversible, and those that are thermodynamically irreversible. A reversible process in thermodynamics is one in which no energy is dissipated (spread) into empty energy states available in a volume, from which it cannot be recovered into more concentrated forms (fewer quantum states), without degradation of even more energy. A reversible process is one in which this sort of dissipation does not happen. For example, conversion of energy from one type of potential field to another, is reversible, as in the pendulum system described above. In processes where heat is generated, quantum states of lower energy, present as possible exitations in fields between atoms, act as a reservoir for part of the energy, from which it cannot be recovered, in order to be converted with 100% efficiency into other forms of energy. In this case, the energy must partly stay as heat, and cannot be completely recovered as usable energy, except at the price of an increase in some other kind of heat-like increase in disorder in quantum states, in the universe (such as an expansion of matter, or a randomization in a crystal).

As the universe evolves in time, more and more of its energy becomes trapped in irreversible states (i.e., as heat or other kinds of increases in disorder). This has been referred to as the inevitable thermodynamic heat death of the universe. In this heat death the energy of the universe does not change, but the fraction of energy which is available to do produce work through a heat engine, or be transformed to other usable forms of energy (through the use of generators attached to heat engines), grows less and less.