Historical Burdens on Physics

edition 2022

FRIEDRICH HERRMANN AND GEORG JOB (ED.)

Friedrich Herrmann and Georg Job

Historical Burdens on Physics

2022

Illustrations: *Friedrich Herrmann*

Translation from German: *Friedrich Herrmann*

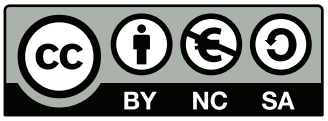


TABLE OF CONTENTS

Introduction

Today's science curriculum is the result of a process of evolution. It reflects the process of the development in great details. Those who are learning science have to follow a path that is similar to the course of the historical development. They have to take detours, to overcome unnecessary obstacles and to reproduce historical errors. They have to learn inappropriate concepts and employ outdated methods. When developing the Karlsruhe Physics Course we have tried to eliminate such obsolete concepts and methods.

In the history of science it happened time and again that important works and results were not accepted by the scientific community: When they arrived it was too late. A change, – although it might have been extremely useful – had become too tedious. Here are three examples:

1. The physical quantity entropy had three chances to become a quantity that would be easy to grasp, even for a beginner; the first chance was after the works of Joseph Black and Sadi Carnot, the second after the work of H. L. Callendar and the third through the book *A new concept of thermodynamics* by Georg Job. All of these chances were missed. The corresponding ideas were incorrectly interpreted or simply ignored.

2. The physical quantity force with the corresponding terminology – a sophisticated construction of Newton – turned out to be the strength of the current of momentum. The corresponding publication from 1908 by Max Planck remained virtually unnoticed.

3. The first 50 years after the introduction of the energy into physics it was not clear if energy obeys a local conservation principle. It was expected but not proven. For that reason a terminology came into use that took these doubts into account. The publication of 1898 by Gustav Mie, in which it is shown that energy obeys a continuity equation did not lead to a more appropriate and simple language. We still speak about energy as if we had to be prepared that one day actions at a distance might be discovered.

In a certain sense, the growth of scientific knowledge is similar to the evolution of biological systems. Every person who is teaching science acquired his scientific knowledge before. Thus, facts are first received and later transmitted. This transmission, however, doesn't proceed without changes, because research brings new results and the teaching person will try to take these results into account. Such changes can be compared with mutations in genetics.

Generally, the changes and improvements a teacher makes concern only his specialty, whereas the general structure of science will be transmitted without alterations. Thus, the basic knowledge is not subject to the same selective pressure as more recent developments. Accordingly, the new knowledge is essentially attached to the old one without questioning the old nucleus. In the theory of evolution this phenomenon is known as *prolongation*. A greater restructuring will be more and more difficult, whereas the driving force for such changes becomes weaker and weaker. In other words: The more complex a system is the more conservative it will be. For this reason, the scientific knowledge reflects quite accurately its historical development. This statement reminds us of a rule which every student of biology has to learn: E. Haeckel's biogenetic law according to which "ontogeny recapitulates phylogeny".

As a result, detours in the development of scientific knowledge may be preserved. Constructions which, in a larger context, reveal to be superfluous or inappropriate may be maintained. An old transient state may survive as a *living fossil* as geneticists like to call such a phenomenon. Even apparent errors may survive. Considering the actual physics syllabus very much can be learned about the history of physics. Indeed, one can even pursue a kind of archaeology in this manner. As a consequence, every student has to reproduce the historical developments. The individual student's process of learning proceeds, often up to the details, according to the same pattern as the development of science as a whole.

By citing the analogy between the evolution of science and that of biological systems, we want to show that the development of science toward more and more inflexibility is an inevitable and normal process and it is not a daring accusation to say that science is unnecessarily complicated and cumbersome. When we claim that science, as a whole, is in a bad state we don't mean that scientists have been incompetent. Those who worked for the advancement of science usually did the right thing in their time. Just like a biological fossil in a remote time accomplished an important function, many components of science, which nowadays may be considered to be superfluous or inappropriate, have played an indispensable part in the past.

For many years, we have been searching systematically for subjects in the physics syllabus which might be considered historical burdens, i.e. superfluous or inappropriately presented subjects. Since 1994 they appeared regularly as a column in the school science review *Praxis der Naturwissenschaften*.

In order to discover such obsolete concepts a certain attitude is necessary which might be considered a lack of respect. Indeed, it is a kind of disrespect in view of convictions which have developed by mere habit and indolence. It is no disrespect, however, for the achievements of the scientists who developed a new concept in the first place.

Each of the articles is structured in the same manner. First, we introduce the *subject*. Then we describe what we believe is the inappropriateness or obsoleteness in the subject: the *deficiencies*. Next, we briefly explain how the subject came into being, i. e., what was the positive role it had played in the past: the *origin*. And finally some comments are made about how to cope with the problem: the *disposal*.

F. Herrmann und G. Job: The historical burden on scientific knowledge, Eur. J. Phys. **17** (1996), S. 159

Friedrich Herrmann, Karlsruhe Institute of Technology
Georg Job, University of Hamburg

Introduction

Today's science curriculum is the result of a process of evolution. It reflects the process of the development in great details. Those who are learning science have to follow a path that is similar to the course of the historical development. They have to take detours, to overcome unnecessary obstacles and to reproduce historical errors. They have to learn inappropriate concepts and employ outdated methods. When developing the Karlsruhe Physics Course we have tried to eliminate such obsolete concepts and methods.

In the history of science it happened time and again that important works and results were not accepted by the scientific community: When they arrived it was too late. A change, – although it might have been extremely useful – had become too tedious. Here are three examples:

1. The physical quantity entropy had three chances to become a quantity that would be easy to grasp, even for a beginner; the first chance was after the works of Joseph Black and Sadi Carnot, the second after the work of H. L. Callendar and the third through the book *A new concept of thermodynamics* by Georg Job. All of these chances were missed. The corresponding ideas were incorrectly interpreted or simply ignored.

2. The physical quantity force with the corresponding terminology – a sophisticated construction of Newton – turned out to be the strength of the current of momentum. The corresponding publication from 1908 by Max Planck remained virtually unnoticed.

3. The first 50 years after the introduction of the energy into physics it was not clear if energy obeys a local conservation principle. It was expected but not proven. For that reason a terminology came into use that took these doubts into account. The publication of 1898 by Gustav Mie, in which it is shown that energy obeys a continuity equation did not lead to a more appropriate and simple language. We still speak about energy as if we had to be prepared that one day actions at a distance might be discovered.

In a certain sense, the growth of scientific knowledge is similar to the evolution of biological systems. Every person who is teaching science acquired his scientific knowledge before. Thus, facts are first received and later transmitted. This transmission, however, doesn't proceed without changes, because research brings new results and the teaching person will try to take these results into account. Such changes can be compared with mutations in genetics.

Generally, the changes and improvements a teacher makes concern only his specialty, whereas the general structure of science will be transmitted without alterations. Thus, the basic knowledge is not subject to the same selective pressure as more recent developments. Accordingly, the new knowledge is essentially attached to the old one without questioning the old nucleus. In the theory of evolution this phenomenon is known as *prolongation*. A greater restructuring will be more and more difficult, whereas the driving force for such changes becomes weaker and weaker. In other words: The more complex a system is the more conservative it will be. For this reason, the scientific knowledge reflects quite accurately its historical development. This statement reminds us of a rule which every student of biology has to learn: E. Haeckel's biogenetic law according to which "ontogeny recapitulates phylogeny".

As a result, detours in the development of scientific knowledge may be preserved. Constructions which, in a larger context, reveal to be superfluous or inappropriate may be maintained. An old transient state may survive as a *living fossil* as geneticists like to call such a phenomenon. Even apparent errors may survive. Considering the actual physics syllabus very much can be learned about the history of physics. Indeed, one can even pursue a kind of archaeology in this manner. As a consequence, every student has to reproduce the historical developments. The individual student's process of learning proceeds, often up to the details, according to the same pattern as the development of science as a whole.

By citing the analogy between the evolution of science and that of biological systems, we want to show that the development of science toward more and more inflexibility is an inevitable and normal process and it is not a daring accusation to say that science is unnecessarily complicated and cumbersome. When we claim that science, as a whole, is in a bad state we don't mean that scientists have been incompetent. Those who worked for the advancement of science usually did the right thing in their time. Just like a biological fossil in a remote time accomplished an important function, many components of science, which nowadays may be considered to be superfluous or inappropriate, have played an indispensable part in the past.

For many years, we have been searching systematically for subjects in the physics syllabus which might be considered historical burdens, i.e. superfluous or inappropriately presented subjects. Since 1994 they appeared regularly as a column in the school science review *Praxis der Naturwissenschaften*.

In order to discover such obsolete concepts a certain attitude is necessary which might be considered a lack of respect. Indeed, it is a kind of disrespect in view of convictions which have developed by mere habit and indolence. It is no disrespect, however, for the achievements of the scientists who developed a new concept in the first place.

Each of the articles is structured in the same manner. First, we introduce the *subject*. Then we describe what we believe is the inappropriateness or obsoleteness in the subject: the *deficiencies*. Next, we briefly explain how the subject came into being, i. e., what was the positive role it had played in the past: the *origin*. And finally some comments are made about how to cope with the problem: the *disposal*.

F. Herrmann und G. Job: The historical burden on scientific knowledge, Eur. J. Phys. **17** (1996), S. 159

Friedrich Herrmann, Karlsruhe Institute of Technology
Georg Job, University of Hamburg

1

General Subjects

1.1 Technical terminology

Subject:

"Technical term…: a well-defined, special designation for a well-defined concept in a particular technical field." [1]

"Technical terminology differs from everyday language among other things, in that its concepts are unambiguously denominated…" [2]

Deficiencies:

Technical terminology is considered an exact language. When we know to which technical field a statement belongs, the statement is unambiguous – this at least is the widely held view. Probably it is the view of non-specialists. Specialists know or should know that this appraisal is not true.

As an example we consider the word *force* and the various concepts that were designated by this word or its latin equivalent *vis*. It is well-known that in the 17th and 18th century the word covered various concepts. Some authors used it for what we call today momentum, others used it for what we now call kinetic energy (*vis viva*), but also what we still today call a force, namely the quantity ***F***. One might believe that this ambiguity was due to the striving for clarity that was still going on. However the use of the word within physics was by far not consistent at the time when clarity about the underlying physics was reached. The following citation stems from a text book from 1912: "We call the product of half of the mass with the square of the velocity of the moved weight its living force." [3] And even today in physics the word is often used for the quantity energy [4]. But in addition, a new competitor appeared in the arena. The thermodynamics of irreversible processes took up the word for its purposes, i.e. to describe the "drive" or the "cause" of any dissipative transport process: "We have seen in the preceding section, that for the appearance of an irreversible phenomenon there exist a series of causes: for instance a temperature gradient, a concentration gradient, a potential gradient or a chemical affinity. In the thermodynamics of irreversible processes these quantities are called 'forces'…" [5]. Moreover, the term "electromotive *force*" has survived undisputed until this day.

One might believe that only our ancestors were able of such an irresponsible handling of the scientific language. But that is not the case. Just now we can observe that the innocent word "force" is being engrossed by a new group of specialists: the particle physicists. It is not easy to understand what exactly a particle physicist means when he speaks about a force. Apparently they use the terms "force" and "interaction" synonymously [6]: "Two of the three interaction particles of the weak force are electrically charged. Therefore they are subject to the electromagnetic force. Thus, they can emit photons and attract one another." Apparently, in this context the word "force" is not used as a name of a physical quantity.

With some attention many other examples of such a change of meaning of a scientific term can be detected.

The *bit* was introduced as a measuring unit of Shannon's amount of data. But later it was used synonymously for "two-state quantum system". The upgrade of the term reached a new level, when the term qubit appeared.

The term "orbital" was coined as a name for a concept that had to replace the "trajectory" concept that was banished by quantum mechanics. Later its meaning was transferred to two more physical concepts. For some it designates a one-particle wave function [7], for others the square of the wave function [8].

In spite of DIN and ISO, SI and IUPAP, technical terms are not used with a unique meaning. The scientific language is not fundamentally different from the colloquial language. Both of them undergo a continuous development. This process is essential for the colloquial language. In linguistics this phenomenon is called a *semantic change*. For the scientific speach such changes are the cause of misunderstandings and learning difficulties. Problems can arise when the user of the language, and in particular the teacher are not aware of the ambiguity of a scientific term.

Origin:

The scientific language is subject to the same linguistic laws as the colloquial language. It is in a continuous process of change and development. In both areas new meanings often appear due to an insouciant handling of the language. Whenever a new scientific special subject emerges, scientists begin to speak a slang, which at the beginning is not meant to be definite, but which finally condenses into what later is considered the technical language of the new area.

Disposal:

As a teacher, do not take part in every quirk of the representatives of a scientific or technical speciality. Do not, without good cause, use a scientific term in various meanings. For instance: Distinguish between the two meanings of the word field: 1. as a name for a physical quantity and 2. as a name for a physical system. If a word is firmly established with two meanings, and both of them are indispensable, advise the students of the problem.

[1] Duden, Deutsches Universalwörterbuch (German Universal Dictionary), Dudenverlag Mannheim 1989, keyword: Fachausdruck (technical term).

[2] Wikipedia, December 2006, keyword: Fachsprache (technical language).

[3] *E. Riecke*: Lehrbuch der Physik, Verlag von Veit & Comp. Leipzig, 1912, p. 63.

[4] *F. Herrmann*: Force and energy, article 5.15

[5] *S. R. de Groot*: Thermodynamik irreversibler Prozesse, Bibliographisches Institut Mannheim 1960, p.4.

[6] DESYs KworkQuark 2006 <http://www.kworkquark.net/>

[7] *K. Bethge and G. Gruber*: Physik der Atome und Moleküle, VCH Weinheim 1990, p. 199: "In the chemical literature one-particle wavefunctions are called orbitals…"

[8] dtv-Atlas zur Chemie, dtv München 1983, p. 23: "Instead the term orbital designates the probability of finding an electron (electron density distribution) within an atom."

Friedrich Herrmann

1.2 Interaction

Subject:

The term “interaction” is used in physics in different contexts. Thereby its meaning does not always coincide with that in the colloquial language.

Deficiencies:

In physics the term “interaction” stands for several different phenomena and processes.

1. The word is used when two bodies exert forces on each other in the sense of Newton’s Third Law. At first glance it seems that here the term interaction is appropriate. When a body A exerts a force on another body B, according to the Third Law B also exerts a force on A. Since we say that a force “acts”, we are dealing here with two actions: A on B and B on A. We thus have an “interaction”, even in the colloquial use of the word. However, this observation leads us to a first problem. The term interaction is suitable only as long as one describes the process with the Newtonian model of an action at a distance. The description refers to only two systems, which are well separated: body A and body B. However, since more than a hundred years we no longer need this provisional description, since we are now convinced that any action is based on the transport of a physical quantity. In particular, a Newtonian force is nothing else than a momentum transport. If the spring (we imagine it to be massless) pulls bodies A and B toward each other, Fig. 1, the momentum of A increases and that of B decreases. But it is not that at B momentum disappears and at A reappears. Rather, it is transferred via an intermediate medium or system C – in our case the spring. Thus it is possible to specify how the momentum gets from B to A. On these grounds it is not appropriate to say that there is an interaction. When a partner B gives something away and A receives it, it would be more convenient to say that there is a transfer, a transport or a transmission. When someone is pouring water from one bucket into another, it would not characterize the process suitably to say that there is an interaction between the buckets.

Fig. 1. The spring is under tensional stress. The momentum of A increases, that of B decreases (the “negative momentum” of B increases).

2. In particle physics one distinguishes between the particles of matter (hadrons and leptons) and the bosonic interaction particles (sometimes called interaction carriers or force mediators). In this field the term interaction means that a certain particle is created or annihilated. Since there are four kinds of boson fields there are also four different interactions: electromagnetic, gravitational, weak and strong. These processes include the interaction in the classical sense, i.e. the case that between two particles (of matter) momentum is transferred, while the nature of the particles is not changed (an example is electron-electron scattering). In addition, they include processes in which two interaction particles “interact” (example: photon-photon or gluon-gluon). But there are also processes in which particles of matter change their nature (an example is the beta decay in which a proton transforms into a neutron, an electron and a antineutrino). It can be seen that here the term “interaction” does no longer coincide with the colloquial interaction. Rather the term describes something that might better be called a reaction (in the sense of chemistry).

3. In other subfields of physics the word is used in an even broader sense, namely for the description of the various processes in which two or more subsystems are involved. Now, one can hardly conceive a process for which this is not the case, so that eventually everything becomes an interaction. It sounds all well and scientific, when something is called an interaction, even if nothing concrete is said.

Origin:

Newton did not use the term, there was no “interactio”. For him there were only “actio” and “reactio”. In the following time his Third Law was not yet called law of interaction, but law of counteraction. By the end of the 19th century, the term “interaction” appears in the scientific literature, see for example in the *Science of Mechanics* by Ernst Mach. But the word got its immense popularity much later, probably in the second half of the 20th century, when every physical process involving two sub-systems became an interaction.

Disposal:

1. Formulate Newton’s Third Law without actions at a distance: The momentum which body B loses, is transferred to body A.

2. In the context of the four bosonic fields the term has acquired a specific meaning. Although the word was not the best choice, we have to accept it as a new technical term with its own meaning.

3. A parsimonious use of the word makes every text clearer.

Friedrich Herrmann

1.3 Quartz clock and Geiger counter

Subject:

From physics text books students learn the working principle of the Geiger counter but not that of the clock. They learn the details about the liquid-in-glass-thermometer but nothing about the thermocouple. They learn about the electronic processes going on in a laser but not those in an incandescent lamp or the flame of a candle. At an advanced level they study the mass spectrometer and the Wien filter, but not the Fourier spectrometer. Isothermal processes are discussed quantitatively, but isentropic processes not at all. At University students learn why the sky is blue, but not why the rest of the world is red, green, grey, black or white.

Deficiencies:

When we design a curriculum or establish educational standards, but also when we prepare our own lessons or lectures, we must make informed choices. Which physical quantities do we introduce and which do we leave out? How many teaching hours do we schedule for mechanics, how many for thermodynamics and how many for electricity? For which phenomena and processes shall we give a microscopic interpretation and which do we treat on a macroscopic scale? What meters or sensors do we introduce in the classroom?

When looking at our curricula and textbooks, we may notice that often the choice was not a good one. This is shown by the foregoing examples.

Origin:

Often a topic gets into the teaching repertoire by fortuity. Then it becomes a custom and its entitlement is no longer questioned. In addition, a tradition of "examination problems" has developed that assures the survival of certain topics. Another element that stabilizes certain particular teaching subjects is the equipment produced by the teaching-materials companies. The inertia of the whole system consisting of teachers, professors, teacher training institutions, University and secondary school text books, their authors, examination habits, and the equipment for demonstration and lab experiments is very large.

Disposal:

To select the topics for curricula, textbooks or for our own classes, we recommend the following method: First chose any subject, that might be considered a candidate for the curriculum. Next try to find "competitors", i.e. topics that can be considered equivalent in any respect: their level of difficulty, their usefulness for applications, their value as a subject of general education, etc. The competitors may be subjects which are normally not found in the curriculum. The candidate survives only if we find good reasons that it is more important than the competitors, that currently are not in the curriculum.

We use this method, because it is not enough to give reasons why an issue is important. Such reasons can be found for any subject, and in general it is easy to present them convincingly. Thus, the important thing is, that a subject has to win against its competitors.

Let us consider an example: The initial proposal is to introduce the electric field strength. Which are the competitors? There are several kinds of them. First, there is the other vector quantity that allows to describe an electric field, the electric displacement. Next, there is a scalar quantity that allows to describe the electric field: the electric potential. Other competitors are analogue quantities for the magnetic field and the gravitational field. Now the following questions must be answered: If we introduce the electric field strength, why not the magnetic and the gravitational field strength? Or: If we do not introduce the gravitational field strength, why then should we introduce the electric field strength?

We will not find it difficult to answer the questions in this case. The following situation is more difficult. The initial subject that is proposed is the thermal dilatation of liquids and solids. Again we look for competitors, and these will mainly be other physical properties of materials: thermal, mechanical, electric, magnetic and optic. We will compare their order of magnitude. (The thermal expansion of liquids and solids is of the order of 10^{-3}.) We will also compare with regard of their importance for a general physical understanding and with regard to technical applications. In this case the number and importance of the competitors is so large, that the thermal expansion of liquids and solids hardly survives.

A similar conclusion is unavoidable regarding the Geiger counter. Competitors are numerous instruments some of which are as exotic as the Geiger counter, but others are as omnipresent as the quartz clock or the CCD matrix of a digital camera.

Among our examples in the section "subject" there are topics that currently are treated at school and that would not survive such a process, and there are other topics that unjustifiably are not part of the curriculum.

Friedrich Herrmann

1.4 Measuring precision

Subject:

"A measurement is the empirical determination of the actual value of a physical quantity."

Deficiencies:

Measuring the value of a physical quantity is a standard task in physics. Measurements are carried out in order to find out or to verify the relation between physical quantities. When explaining why a measurement is necessary one often suggests the following idea: Before making the measurement the value is unknown, after the measurement it is known. Thus there are two states or situations: "not measured" and "measured". Our citation is an extreme example for such a point of view. It says in addition that there exists an actual value. Sometimes it is stressed that we have to make a measurement because our senses are imprecise and unreliable.

This view is unfortunate in two respects.

First: It is not true that before executing a measurement nothing is known about the value of the physical quantity in question. And second: It is not the case that after the measurement we know the actual or exact value. Before making the measurement we know that the value is situated in a certain interval, which may be very large; after the measurement we also know that the value is in a certain interval, but this interval is smaller than that before the measurement. If by doing the measurement the interval has been strongly reduced, then it is a good measurement. If it is only slightly reduced the measurement is not so good.

Based on this observation we can define a number which characterizes the quality of a measurement of the quantity X: the ratio between the interval before and that after the measurement

$$\frac{X_{b2} - X_{b1}}{X_{a2} - X_{a1}}.$$

The index b refers to "before" and the index a to "after". A more convenient definition would be the binary logarithm (lb) of this ratio

$$M = \mathrm{lb}\left(\frac{X_{b2} - X_{b1}}{X_{a2} - X_{a1}}\right)\mathrm{bit} \qquad (1)$$

since it represents the information gain achieved by the measurement. It tells us by how many bits the information content of the value of a physical quantity has increased by the measurement. Suppose that before carrying out the measurement it is known that the value of the quantity under consideration is situated between 10 and 12, and after the measurement we know it to be between 10,6234 and 10,6236. We calculate

$$M = \mathrm{lb}\left(\frac{12 - 10}{10{,}6236 - 10{,}6234}\right)\mathrm{bit} = 13{,}3\,\mathrm{bit}$$

The measuring instrument has provided 13,3 bit*.

Origin:

In school physics it is common to classify a measurement as good if the precision is better than about 5%. It is considered bad if the precision is worse than 20 % more or less. This appraisement is rather arbitrary. Probably it is due to the fact that the old pointer instruments had a measuring precision of around a few percent. It may also be related to the fact that we can determine the values of several quantities, like distances, velocities and masses, by using our senses with a precision around 10 % to 50 %. The idea might have been that an operation is called a measurement only if it supplies values that are more precise than those which we get by using our senses.

Disposal:

We recommend to take a measuring result seriously even if the precision is in a range that usually is considered as imprecise. A "measurement" that is realized with our senses is not necessarily a bad measurement, i.e. the information increase M can be important.

*The definition is reasonable only as long as the uncertainty is small compared with the measured value. However, it can be generalized in such a way that this case is also covered:

$$M = \mathrm{lb}\left(\frac{\lg\frac{X_{v2}}{X_{v1}}}{\lg\frac{X_{n2}}{X_{n1}}}\right)\mathrm{bit}$$

Suppose it is known that the number of protons in the universe is between 10^{70} and 10^{90}. Now somebody is able to show by means of some astrophysical measurement, that the value is situated between 10^{75} and 10^{85}. Our formula tells us, that the information gain is

$$M = \mathrm{lb}\left(\frac{\lg 10^{20}}{\lg 10^{10}}\right)\mathrm{bit} = \mathrm{lb}\ 2\,\mathrm{bit} = 1\ \mathrm{bit}$$

In the case that the precision is small in comparison with the measured value, the equation simplifies to equation (1).

Friedrich Herrmann

1.5 Linear characteristics

Subject:

In the course of the lessons of mechanics and electricity the students get acquainted with the following linear relationships:

	Equation	*Name*
(1)	$F = -D \cdot s$	Hooke's law
(2)	$p = m \cdot v$	none
(3)	$F = k \cdot v$	sometimes Stokes' law of friction
(4)	$n\Phi = L \cdot I$	none
(5)	$Q = C \cdot U$	none
(6)	$U = R \cdot I$	Ohm's law

Deficiencies:

The equations are part of a common structure of mechanics and electricity. They describe for each of both disciplines three passive linear components.

Each of the six equations is valid only within a sufficiently small range of the pertinent independent variable. A spring does no longer obey Hooke's law if it is overstretched. Momentum is no longer proportional to velocity if the velocity is no longer small compared with *c*. The frictional force is no longer proportional to the velocity if turbulence sets in. Magnetic flux and electric current intensity are no longer proportional to one another if the solenoid deforms under the action of the magnetic field. Electric charge and voltage do not obey a linear relationship if the distance of the capacitor's plates changes under the influence of the tensional stress within the electric field. A resistor does no longer conform to Ohm's law if the electric current gets too strong.

It is seen, that the linearity is each time a special case. This special case, however, is particularly important, since it is always valid provided that the independent variable's value is not too great.

The well-known examples of the mechanical and the electric harmonic oscillator show how the equations are interrelated. In each of the two differential equations for damped harmonic oscillations three of the components, that correspond to the equations (1) to (6) are represented by a summand. To each component of the mechanical oscillator there is a corresponding component in the electric oscillating circuit. Due to the similarity of the mathematical structure of the differential equations the solutions of these equations have also the same structure.

Seen in this way, a relationship between the equations (1) to (6) becomes apparent and it would be logical to teach this structure to our students. Actually, we are used to proceed quite differently.

First, there are the names: We have well-established names only for equations (1) and (6): Hooke's law and Ohm's law. This observation is not at all marginal. An equation with a name is perceived as more important than a nameless formula.

More important is how the equations are "sold" to the students: Only equations (1), (5) and (6) are introduced as described above, i.e. as the expression of an observable linearity and as the definition of the factor of proportionality.

Relation (2) is presented as the equation that defines momentum. Therefore it does not reflect any observable property. As a pure definition it is not a law of nature. From this point of view it seems natural that the equation has no name.

Equation (3) is that law of mechanical friction which corresponds to Ohm's law in electricity. The students learn it, if at all, only peripherally. In mechanics it is usually not mentioned. Apparently, friction between solids bodies is considered more important. But it is treated as the typical mechanism of friction in the context of oscillations. (It is obvious why.) Moreover, the law is used when teaching the Millikan experiment. One must hope that the students will not believe that Stokes' friction is a peculiarity of the Millikan experiment. The shock absorber of a car, which is not less important than springs and brakes, is usually not treated at school.

Origin:

The six equations have been discovered over a time span of about 200 years by different persons in different contexts. Although it is not difficult to recognize the structure, and although this structure is worth a whole semester's lecture at Faculties of Engineering, the physics curriculum has never taken notice of it, maybe due to the pronounced sense of tradition of the physicists.

Disposal:

It would be completely unpromising to try to remove a name from an equation that is attached to it since more that a hundreds years or to give a name to an equation that did not have one in the past. (Although it would not be unreasonable to call equation (2) Huygens' law or Descartes's law, in honor of one of its discoverers.) All we can do is to show and to emphasize the analogy and to address the questions of the asymmetric treatment in the text books.

We also show that the proportionality between p and v (equation (2)) is indeed observable. Then the inertial mass is defined as the corresponding factor of proportionality. Together with Newton's second law $dp/dt = F$ we get the beloved (too beloved?) relation: $F = m \cdot a$.

Friedrich Herrmann

1.6 Integrals of motion

Subject:

In theoretical mechanics, integrals of motion play an important role: quantities whose values remain constant in time. A system with n degrees of freedom has $2n - 1$ such integrals. One often calls these quantities conserved quantities:

(1) "A function $f(q,\dot{q},t)$ is called a conserved quantity or integral of motion, if

$$\frac{df(q,\dot{q},t)}{dt} = 0$$

or $f(q,\dot{q},t)$ = constant, holds for all trajectories $q_i(t)$, that fulfill the *Lagrange*-equations [1]."

(2) "Apparently, the momentum is a conserved quantity, if its temporal derivative disappears, i.e. if the forces $\vec{K}_1$ and $\vec{K}_2$ are equal and opposite during the entire course of the motion, if thus

$\vec{K}_1 + \vec{K}_2 = 0$." [2]

Deficiencies:

In theoretical or analytical mechanics the expression "conserved quantity" has a different meaning from that in other fields of physics.

In general, i.e. if we refrain from analytical mechanics, one uses the designations "conserved" or "not conserved" in order to characterize a substance-like physical quantity. (A quantity is substance-like if a density and a current can be attributed to it.) Some substance-like quantities are conserved, like energy, momentum and electrical charge, and others are not, for instance entropy. Conservation or non-conservation, respectively, is a universal property of a quantity. It is not the characteristic of a certain function, a certain system or a certain process. It also makes no sense to speak of the conservation or non-conservation of a non-substance-like quantity. Temperature, for instance, is neither conserved nor non-conserved.

In theoretical mechanics, on the contrary, the word "conserved quantity" stands for "integral of the motion" (see our first quotation). An integral of motion is not necessarily substance-like and is often not intuitive. An example is the *Runge-Lenz* vector. The *Runge-Lenz* vector is time-independent for the *Kepler* problem. According to the usage of theoretical mechanics it is a conserved quantity of the *Kepler* problem. However, the *Runge-Lenz* vector is not a substance-like quantity, because no density and no current density can be defined for it. In addition, it is not always time-independent, but only in the *Kepler* problem.

According to the practice of theoretical mechanics the quantities energy, momentum and angular momentum are sometimes conserved and sometimes not (see our second quotation).

Origin:

Theoretical mechanics is one of the most elegant physical theories. It also plays an important role as a basis of other theories: It requires only few modifications to become quantum theory. This perfection may be due to the fact that it was completed quite independently of other fields of physics. Thereby it has developed its own vocabulary. Among other things the designation "conservation" is used in a different sense than elsewhere. This cannot always be noticed because in some cases the meaning overlaps with that in the other areas of physics. This use may also be the cause of a somewhat unfortunate formulation of the true and universal conservation of a quantity. Instead of characterizing a conserved quantity by saying that it cannot be produced or destroyed, it is said that the value of the quantity is constant in a closed system.

Disposal:

One distinguishes between the concepts "integral of motion" and "conserved quantity", as it is done, for instance, in Landau-Lifshitz [3]: "Among them [the integrals of motion] are some whose constancy has a deeper cause, that is related to the basic properties of time and space – its homogeneity and its isotropy. All these so-called conserved quantities have in common the important property of being additive."

[1] *F. Kuypers*: Klassische Mechanik. Physik-Verlag, Weinheim, 1983, S. 38.

[2] *W. Macke*: Mechanik der Teilchen. Akademische Verlagsgesellschaft, Leipzig, 1962, S. 240.

[3] *L. D. Landau* and *E. M. Lifschitz*: Theoretische Physik kurzgefaßt I. Akademie-Verlag, Berlin, 1973, S. 17.

Friedrich Herrmann

1.7 Conservation laws

Subject:

It is possible to state whether any extensive quantity is conserved or not. Some extensive quantities obey (as far as we know) a universal conservation law: energy, momentum, angular momentum, electrical charge, lepton number, baryon number, color charge, etc. There is one quantity that obeys a “half conservation law”: Entropy can be produced, but not destroyed. Each quantity that is not generally conserved may be conserved under certain circumstances. As an example, entropy in reversible processes behaves like a conserved quantity. The amount of substance is not generally conserved, however there are many processes in which it behaves like a conserved quantity.

Deficiencies:

If the extensive quantities are placed in the foreground, then one gets a representation of physics in which the various sub-fields reveal the same structure. Mechanics, thermodynamics, electricity and chemistry appear as special cases of a uniform structure of concepts. In order to be able to take profit of this structural similarity, it is necessary that the various corresponding physical quantities are treated in an analog manner. Therefore it is recommendable that the conservation or non-conservation of the various extensive quantities is treated in analogous ways, on equal footing. However this is not usually done.

For instance, the conservation of energy is presented as one of the most important principles of the whole of physics. The conservation of momentum is dressed in Newton's laws, such a strangely complicated outfit that this simple statement can no longer be recognized. Completely different again is the electrical charge: Over its conservation not a single word is wasted, since it is usually presupposed as obvious. The simple fact that entropy can be produced but not destroyed is sometimes found in schoolbooks in the small print, and usually in the place where instruction never reaches. The non-conservation of the amount of substance is never formulated as a theorem, nor is the fact that for certain classes of processes the amount of substance is conserved. Instead of formulating and applying the simple and useful conservation laws that are known from nuclear and particle physics, precious teaching time is wasted with the discussion of details of special radiation meters.

Origin:

The statements about conservation or non-conservation of extensive quantities reflect the historical development of physics. If the discovery and formulation of such a statement was difficult and took a long time, or if the validity of the statement was questioned for a long time, then much time will also be reserved for teaching the concept, and the statement will be presented as particularly important. The clearest example of this is energy conservation. One might argue that the principle of energy conservation is so fascinating for us because it forbids something with which one could make a lot of money. This is true. However, it also shows the lack of imagination of the would-be perpetual motion inventors, since they could also make a lot of money by breaking any of the other conservation laws.

On the other hand, if the discovery of a conservation or a non-conservation theorem was quick and easy, and if the statement was historically not doubted, then the theorem is also treated quickly in the classroom, or not at all.

Disposal:

Instruction would win if one:

1) clearly formulated conservation or non-conservation for each extensive quantity;
2) clearly pointed out the importance of conservation or non-conservation (particularly in the case of electrical charge, amount of substance, lepton number and baryon number);
3) did not exaggerate the importance of conservation (as with the energy).

Friedrich Herrmann

1.8 Particles everywhere

Subject:

“Each time that the molecules of the hot water vapor impacts on the receding blade of the turbine wheel, they transfer part of their kinetic energy to it and bounce back with reduced velocity.”

“Due to the great wavelength and the low friction between the water molecules a Tsunami is only barely damped.”

“Since the demand for electricity is often lowest when the wind is blowing strongest, Denmark must sell its electron surplus for a few cents to the neighboring countries.”

Deficiencies:

First citation: Only very few of the water molecules come in touch with the turbine blades. So it is exaggerated to say “*the* molecules”. Only very few molecules would be correct.

Second citation: Molecules are not rough and there is not friction between individual molecules.

Third citation: A person with an education in science will understand what is meant: something different from what is said. (For an alternating current with a typical current density the electrons oscillate back and forth by only several microns. So it is greatly exaggerated to say that the electrons leave Denmark. In addition, the power line consists of a forward and a return line.) Who is not so familiar with physics will believe that electrons really move from Denmark to Germany or Sweden.

One might say that these small blunders are not worth mentioning if they were not symptomatic for a pronounced predisposition of most physicists to explain everything by referring to particles. A physical phenomenon is not really understood as long as it is not reduced to the behavior of particles – this is a widespread opinion.

One speaks about water molecules when water is meant, of photons when light is meant and of electrons when electric charge, or as in our last citation, energy is meant.

Of course, one can tell for any physical process what is going on at the microscopic scale, and there are always particles which do something. However, what they do is not always enlightening for the actual problem. One does not better understand the steam turbine with the water molecules than with the water vapor, one does not better understand the Tsunami with the water molecules than with the water, one does not better understand electrical phenomena with the electrons than with the electric charge.

Physics works with physical quantities and we (and our students) have no problems in dealing with them. It is true that the particles are intuitive but we also can acquire an intuitive idea of physical quantities. We can imagine electric charge and energy as fluids that can flow and for which we can establish a balance. When only relying on the particles in many situations it becomes more difficult to come to an understanding or to a formal mathematical description. In addition, the particle description does not represent a deeper truth.

Out of the intricate interplay of many particles on a higher level emerge new phenomena that can be described by a simpler theory. In the philosophy of science this phenomenon is called *emergence*. When reducing the behavior of a system to the behavior of its component particles, one often explains the simple by the complicated.

Origin:

“Reductionism” is a general trend. In the 19th century it celebrated its great successes and became generally accepted. Only when a phenomenon was reduced to the mechanics of its constituent particles it was deemed to be understood.

Disposal:

In a steam turbine the steam expands. It presses on the turbine blades. Thereby its pressure and temperature decrease, in the same way as the pressure and the temperature of the air that rises up in the atmosphere.

Instead of arguing with the low friction of the molecules of the water that causes a Tsunami it is sufficient to say that the water is thin fluid.

And finally: Denmark does not export electrons but energy.

Friedrich Herrmann

1.9 Aether and vacuum

Subject:

"The aether as a carrier of electromagnetic fields does not exist, the concept is an unnecessary hypothesis." "The idea of an aether […] as a carrier of electromagnetic waves in the vacuum, has been overcome only with the appearance of the Theory of Relativity."

Deficiencies:

A problem cannot be solved by pretending that the subject of consideration does not exist. The problem was the strange behavior of the aether upon a change of the reference system. This behavior became evident in the Michelson-Morley experiment. The existence of the aether was indeed questioned during a certain period of time by certain researchers, and some of them would have liked to ban the concept from physics altogether. Notwithstanding, after its partial banning the aether was admitted again, though under a new name. It was called vacuum. One might think, that now things are in order again, but according to many books and other texts, space remains empty, as our citations show. This can also be seen in many schoolbooks when the field concept is introduced: A field, so it is said, is empty space with certain properties.

Another deficiency, that is more than only a blemish is the new name. Etymologically, the word vacuum expresses the absence of something or of anything. But it is now used to designate the presence of something. But who would employ the good old name aether is considered as someone who has slumbered away the theory of relativity [1]. In no case should Einstein be cited in favor of empty space. In his later publications he clearly pronounced himself in favor of the aether [2].

Origin:

Since the Michelson-Morley experiment had an unexpected outcome, it was clear that a new theory was necessary to replace time-honored mechanics. Disavowing the existence of an aether was only an act of desperation. It could not solve the problem of the outcome of the experiment. With the appearance of the Theory of General Relativity and later of Quantum Electrodynamics the chimera of the empty space disappeared and the aether came back under a new name.

Disposal:

There are situations where it is justified so speak of an empty space, in the same sense as there is nothing to object against speaking of an empty bottle. We understand that there is no more Whiskey in an empty Whiskey bottle. But we also know that this does not mean that there is not something else in the bottle: air and light for instance. Statements about an empty space can cause problems of understanding however, when it is suggested that empty space contains nothing, or that there is "nothingness". We therefore recommend to use the term "empty space" parsimoniously. One should avoid it completely when introducing the concept of field.

[1] *R. B. Laughlin*: A Different Universe – Reinventing Physics from the Bottom Down, Basic Books, New York, 2005:

"The word 'ether' has extremely negative connotations in theoretical physics because of its past association with opposition to relativity. This is unfortunate because, stripped of these connotations, it rather nicely captures the way most physicists actually think about the vacuum. "

[2] *A. Einstein*: Address delivered on May 5th, 1920, in the University of Leyden:

"Recapitulating, we may say that according to the general theory of relativity space is endowed with physical qualities; in this sense, therefore, there exists an ether … According to the general theory of relativity space without ether is unthinkable; for in such space there not only would be no propagation of light, but also no possibility of existence for standards of space and time (measuring-rods and clocks), nor therefore any space-time intervals in the physical sense. But this ether may not be thought of as endowed with the quality characteristic of ponderable media, as consisting of parts which may be tracked through time. The idea of motion may not be applied to it."

Friedrich Herrmann

1.10 Two effects of a force and three effects of an electric current

Subject:

A force can have two effects: acceleration and deformation of a body.

Electric currents can have three effects: thermal, magnetic and chemical.

Deficiencies:

To get a clear idea of these classifications, which are found in schoolbooks, let us compare the two statements. This is not far-fetched since a force is nothing else than a momentum current. Thus, both classifications are about the effects of currents: a momentum current in the first case and an electric current in the second. Such a comparison brings to light some incongruities.

1. Let us begin with the first effect of a force: the acceleration. It can also be expressed in the following way: A force that is acting on a body can change the momentum of the body. The corresponding electric statement would be: An electric current that is flowing into a body (or out of it) can result in a change of the electric charge of the body. This statement is certainly correct. But why is it not mentioned as one of the effects of an electric current? Because it is obvious and trivial. Now, the acceleration effect of a force is just as trivial. When momentum enters a body and does not leave it simultaneously, it inevitably accumulates in the body.

2. We next consider the thermal or heating effect of an electric current. Heat is generated not only by an electric current. Also momentum currents (forces) can produce heat, namely in frictional processes. Why is it not mentioned as an effect of a force (a momentum current)?

3. The enumeration of the effects of currents is far from complete. So there is yet an electric effect of a force (the piezoelectric effect), an optical effect of a force (birefringence), optical and light effects of the electric current (in an LED), a cooling effect of an electric current (in a Peltier element) etc.

In summary it can be said: The cited effects are no characteristic for the respective current. Not all the effects of the two currents are mentioned, and those which are mentioned are not necessarily the most important. In short: Both classifications contain pretty much arbitrariness.

Origin:

Since mechanics has developed independently from electricity, different models and teaching habits have established in the two disciplines. Too much importance is attributed to momentum conservation (in the form of Newton's laws) as compared to the conservation of electric charge. Mechanical friction on the contrary, as compared to "electric friction" (electric resistance), is stigmatized as a phenomenon that only impairs the mechanical activities.

Disposal:

1. Drop the accelerating effect of a force (momentum current) or include the "charging effect" of an electric current. Our choice would be not to include these two phenomena in the list of effects, since in contrast to the other effects both occur only if the current has divergences.

2. If one engages in a classification then the thermal effect should be mentioned for both currents, the electric and the momentum current.

3. It should be clear that the effects represent only a selection.

Friedrich Herrmann

1.11 Are there physical quantities in nature?

Subject:

In the physical literature one can find the concept of a momentum current. The physical quantity force is nothing else than the intensity of a momentum current, the stress tensor is identical with the tensor of the momentum current density. In a report of the German Physical Society on the Karlsruhe Physics Course it is claimed, that momentum currents do not exist in nature [1]. What is true?

Deficiencies:

Both, because:

1. there are no momentum currents in nature;

2. there are momentum currents in the textbooks.

However, these observations are also true for any other physical quantity. A physical quantity is a mathematical variable in a theory, which on its part is an invention of man [2,3].

In Nature there is not only no momentum current but there is also no electric current, force etc..

Electric charge cannot flow; just as a *mass* cannot hang from a spring or a *volume* cannot contain a gas. *Electrons* can flow, a *body* can hang from a spring and a *container* can contain a gas. Electrons have a property that we describe by its electric charge, the body has a property that we describe by its mass, and the gas is in a container that we describe by its volume.

One may object that it is pedantic to argue this way. It would be nice if one were right; it would be nice if everybody was aware that we are employing a model when speaking about currents of mass, electric charge or energy. If one is aware of this fact, there is no objection against saying that in an electric conductor there is an electric current. Every physicist speaks this way, and that is good. But for the same reason there cannot be an objection against the introduction of momentum currents, convective or conductive.

Not all of those who are using the concept of an electric current or an energy current seem to know that they are using a model.

This becomes particularly clear in that part of the DPG report that deals with heat.

The question of what is heat “really”, is not only a question of the authors of the report; it is as old physics. Only when physics entered a more enlightened phase, it became clear that this is the wrong question. But even today the insight has not arrived everywhere.

There was a long dispute about the question whether heat is a substance or the movement of particles, so as if there was no doubt that heat it something, that exists in the real world and that the role of the scientists was only to discover it and to study its properties. So far the misconception.

It disturbed many people that at the end there appeared several physical quantities that claimed to be a measure of what in everyday life as well as in physics and chemistry could be called heat. Some consider this fact a malice of thermodynamics, one of the reasons why thermodynamics seems to be so difficult. So, in physics the official measure of heat is the quantity dQ. To introduce pupils and students to the concept of heat one often takes recourse to the quantity U, called internal energy. For the chemist by contrast the quantity H, called enthalpy, is the magnitude that measures heat.

Origin:

An unenlightened handling of the basic concepts of science.

Disposal:

Make clear from the beginning that physical quantities are invented, constructed, created by man.

Friedrich Herrmann

[1] Expert opinion on the Karlsruhe Physics Course; Commissioned by the German Physical Society; M. Bartelmann, F. Bühler, S. Großmann, W. Herzog, J. Hüfner, R. Lehn, R. Löhken, K. Meier, D. Meschede, P. Reineker, M. Tolan, J. Wambach und W.

Weber;

http://www.physikdidaktik.uni-karlsruhe.de/kpk/Fragen_Kritik/KPK-DPG%20controversy/Expert_opinion_english.pdf

[2] *A. Einstein, L.Infeld:* Die Evolution der Physik, rororo 1956, S. 29:

„Physikalische Begriffe sind freie Schöpfungen des Geistes und ergeben sich nicht etwa, wie man sehr leicht zu glauben geneigt ist, zwangsläufig aus den Verhältnissen in der Außenwelt.“

[3] *Falk, G., Ruppel, W.:* Mechanik, Relativität, Gravitation, Springer-Verlag Berlin 1973, S. 2:

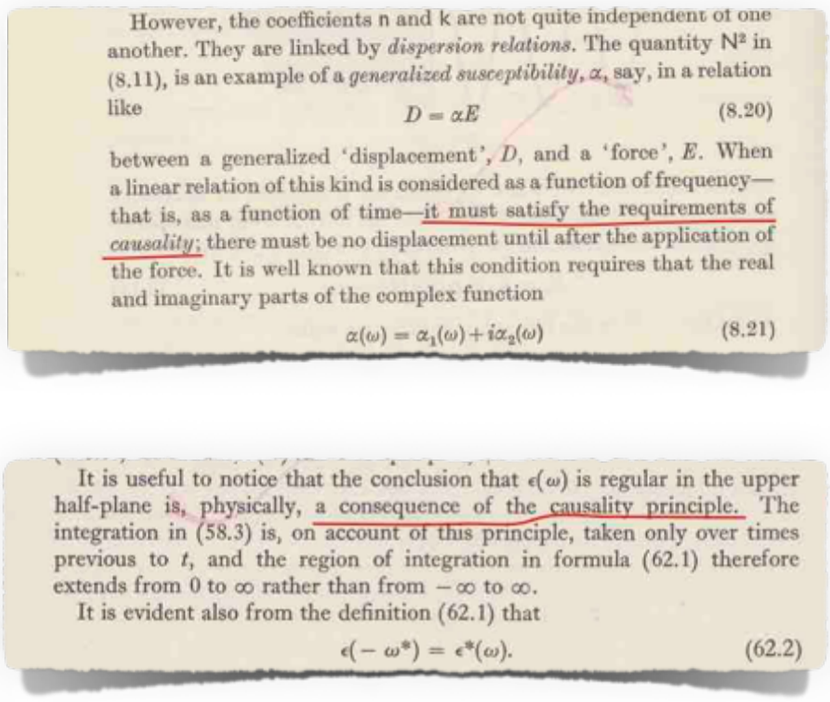
„Schließlich ist es irrtümlich anzunehmen, die Objektivität der Physik bestünde darin, dass ihre Begriffe nichts zu tun hätten mit der menschlichen Fantasie oder überhaupt mit dem Menschen. Tatsächlich sind die physikalischen Größen Erfindungen des menschlichen Geistes, die dazu dienen, die verwirrende Fülle der uns umgebenden Erscheinungen durch durch einfache Regeln überschaubar zu machen.“

1.12 The principle of causality

Subject:

In physics textbooks the principle of causality is usually mentioned in only one context: to justify the Kramers-Kronig relation, i.e. in the context of a very special subject of solid state physics, Fig. 1.

Student teachers will never get in contact with the concept.



However, the coefficients n and k are not quite independent of one another. They are linked by *dispersion relations*. The quantity N^2 in (8.11), is an example of a *generalized susceptibility*, α, say, in a relation like

$$D = \alpha E \qquad (8.20)$$

between a generalized 'displacement', D, and a 'force', E. When a linear relation of this kind is considered as a function of frequency—that is, as a function of time—it must satisfy the requirements of *causality*; there must be no displacement until after the application of the force. It is well known that this condition requires that the real and imaginary parts of the complex function

$$\alpha(\omega) = \alpha_1(\omega) + i\alpha_2(\omega) \qquad (8.21)$$

It is useful to notice that the conclusion that $\epsilon(\omega)$ is regular in the upper half-plane is, physically, a consequence of the causality principle. The integration in (58.3) is, on account of this principle, taken only over times previous to t, and the region of integration in formula (62.1) therefore extends from 0 to ∞ rather than from $-\infty$ to ∞.

It is evident also from the definition (62.1) that

$$\epsilon(-\omega^*) = \epsilon^*(\omega). \qquad (62.2)$$

Fig. 1. Excerpts from two solid state physics text books

Deficiencies:

For those who never had to learn this physical subject there is no problem. If asking such a student for the principle of causality one may get the answer: this is a philosophical subject; it is a matter of course that it is valid; for the physicist there is no problem since in physics its validity is admitted anyway.

A problem may arise for those students who attend the lectures about optical properties of solids, and who also had this opinion. But all of a sudden the hear the lecturer say that for the following calculus the principle of causality is needed. But did he need to study three years in order to come to the point where the principle of causality becomes relevant?

Origin:

Probably only a habit that propagated from generation to generation.

Disposal:

We do not make any objection to the argument in the context of solid state optics. But if one is using such a strong cudgel, one might have mentioned earlier already that one is disposing of it. And if one ever has reflected about the conceptual bases of physics one might have discovered that the causality principle is effective everywhere and all the time.

Friedrich Herrmann

1.13 History of science in the classroom

Subject:

Science is an important part of our culture, and the history of science is worthwhile to be taught at school. Even though we have not the time to give an adequate overview of the history of scientific ideas, we try to sketch the most important developments of science and to introduce the most important protagonists. The names of the researchers that are mentioned or introduced may be considered an indicator for how one is dealing with this subject.

Deficiencies:

To decide which scientists and which of their works are to be discussed at school is a problem and this problem is solved by the schoolbooks more or less satisfactorily.

1. A certain number of researchers has to be mentioned although one would rather not decide to do so: all those who had the luck that an effect, an equation, a rule, a measuring unit, a natural constant or an experiment carries their names. We will not discuss the reasons why one decides to honor a person in this way. But the consequences are problematic for two reasons.

A physical statement appears more important, when it has a proper name and it can be that it is perceived as more important than it actually is. There are cases where an effect or an equation would not be mentioned if it did not carry the name of the researcher who discovered it, such as the Geiger counter, the Wilson chamber, or the Bunsen burner.

We are used to this kind of designations and do no longer ask: Why does the relation $p \sim 1/V$ have a proper name but not the equation $p = m \cdot v$? Why does $F = D \cdot s$ have a name, but not $\Phi = L \cdot I$ or $Q = C \cdot U$? Lenz's law is probably the only law or rule that tells us no more than the sign of a quantity in an equation. But what about the signs in the tens of other equations?

2. When looking for which scientists are mentioned explicitly and purposely one may note that the choice is sometimes arbitrary or imprudent.

This can be seen in particular in the fact that some of the greatest researchers are hardly mentioned at all, as for instance Euler, Descartes, Leibniz or Gibbs.

It is interesting in this context how balance laws are treated: For which quantities can a balance law be formulated and which of these quantities are conserved and which are not conserved?

Energy conservation is usually a subject of discussion and Joule, Mayer, and perhaps Helmholtz are mentioned as the authors.

On the contrary, one learns about how mass became a conserved quantity only if one is a chemistry student. Only then one gets to know Lomonossov and Lavoisier. Physics on the contrary seems to be responsible only for the negation of mass conservation: the mass excess.

It is rather uncommon to learn who brought momentum into physics and who discovered its conservation; thus nothing in this context about Descartes and Huygens. And no word about who introduced angular momentum into physics. Euler is usually mentioned only as a mathematician.

3. Finally one more injustice. Some eminent physicists are mentioned and honored for something that is not their most important achievement. Huygens for instance for his elementary waves (instead for his work on momentum), Daniel Bernoulli for the „Bernoulli equation“ (and not for his contribution to the introduction of angular momentum), Carnot for his 4-step process (and not for the ingenious idea to compare a heat engine with a water wheel).

Origin:

If one examines these examples one may find in each case a different story. But taking all of them together one can conclude: The reason for the often inappropriate choice are coincidence and convention. An equation gets a name if the constellation is appropriate, just like a street gets a name of a person that one wants to honor if the situation is favorable for someone, if he or she has an appropriate lobby. Once the equation or the street has got the name it has got it forever.

Disposal:

I would feel uncomfortable to give a recommendation. I am convinced that the history of physics needs an overhaul, in particular in view of the teaching at school. Regarding the question of which scientists should be mentioned I recommend reservation on the one hand and occupation with the history of physical ideas on the other. And why not a compulsory course about the history of science for teacher students?

Friedrich Herrmann

1.14 When a force acts on the charge of a mass, its momentum changes

Subject:

“A mass is hanging on a spring“, “a charge is accelerated in an electric field”, “a filter transmits specific wavelengths“. In these sentences, which every physicist understands perfectly, the name of an object or a physical system is replaced by that of a physical quantity. What is meant is: “A *body* is hanging on a spring“, “a *charged particle* is accelerated in an electric field”, “a filter transmits *light* of specific wavelengths“.

Deficiencies:

1. Physical quantities are variables in the sense of mathematics. They cannot hang, they cannot be accelerated and they cannot be transmitted by a filter. If, as in the present case, one refers to a physical quantity where an object or a particle is meant, we have to do with what in linguistics is called a metonymy. In colloquial language metonymies are wide-spread. For instance one might say: “The White House has announced ...” where is meant: “The press officer of the White House has announced…”

In the case of our quotations an object is replaced by one of the physical quantities that can be used to describe it: that physical quantity which matters in the corresponding context. Only the mass of the body that hangs on the spring is relevant if one is interested in the oscillation, not its temperature or its color; only the wavelength of the light matters if we want to describe the action of the filter …

2. Identifying an object and a measure of one of its properties is particularly common in the case of mass and electric charge. One would not say that an energy, an entropy or a momentum hangs on a spring. It has to be a quantity that we consider as characteristic for the body. Thus we say that a charge is accelerated only if we have to do with particles with a charge that is characteristic for the particle, such as an electron or a proton. If we accelerate a macroscopic charged body we would rather say the body is accelerated.

3. If the name of a physical quantity is such that it clearly expresses the fact that it is a measure of something, the body will not be denoted by the name of the quantity. Thus one does not say: The body hangs on the spring constant, but on the spring, although in the context only the spring constant matters. The word “constant” is against the identification. One says “Like charges repel each other” but not “Like magnetic pole strengths repel each other” but rather „Like poles repel each other”. The word strength is against the identification of the object and the measure.

4. Sometimes one goes even further: When talking about “the momentum of a mass”, “the volume of a mass” or “the potential of a charge”. Here again, we can notice the special treatment of mass and electric charge. Probably nobody would talk about “the momentum of the energy” or “the temperature of the entropy”, and even less about “the temperature of the pressure” or “the duration of the length”.

But what is the problem with this practice? Often, there is no problem, sometimes there is a small problem and sometimes a big one, which however usually is not noticed. A problem arises for instance when the name “electric field” for a physical system and the designation “electric field strength” for a physical quantity are interchanged or identified, or when a magnetically charged particle (that does not exist in nature) is identified with the physical quantity magnetic charge, or when an electron is considered as electric charge or a photon as energy [1].

Origin:

1. Metonymies are among our common speaking tools. Usually they do not cause misunderstandings. Rather they enrich the language.

2. In physics one aims at more rigor than in colloquial speech; this at least is the self-concept of the physicists. Actually the conceptual rigor often is not worth much. Often in physics a jargon is spoken that is rather serviceable for the mutual understanding of physicists among themselves. However, it is frequent that a misunderstanding or even a scientific dispute comes simply form the improper use of the scientific terms.

3. Historically, the role of the extensive quantities mass, electric charge and entropy was recognized only a long time after their introduction. When it became clear that a phenomenon or a process could be described by means of an extensive quantity, it was at first supposed that one had to do with a kind of substance. So, electric charge seemed to be more than a mathematical tool to describe electric processes. It was believed that one had to do with two kinds of electric fluids. In the same way magnetic fluida and a heat fluidum were supposed to exist. Still today one deals with the physical quantity mass as if was synonymous to matter. In this spirit an extensive quantity was a measure of the amount of something that exists in nature, and it was not distinguished between this fluid and the measure of it. The electric fluid had only one single property, namely that which is described by the electric charge, the heat fluid had only one single property, namely that described by the quantity “heat” (or chaleur”).

Disposal:

The identification of an object and a physical quantity is particularly pronounced among specialist of a certain subject area.

As a school teacher we should think twice whether it is worthwhile to make the mass hanging on the spring. Why not say: “the body hangs on the spring”? Why not say “the charged particle is accelerated” instead of “the charge is accelerated”? The additional effort is small, but the conceptual clearness is great.

One should avoid particularly to say something like: “the energy of the mass” or “the force exerted on the charge” or “the momentum of the mass”.

The two extreme cases are pedantry on one side and gobbledygook on the other. We propose to let the gobbledygook for the specialists, and to aim at conceptual clarity at school, even though it takes some greater effort to say “the current strength of the charge current is 2 A”, instead of “the current is 2 A”. Or “a body with a mass of 2 kg hangs on the spring” “instead of “a mass of 2 kg hangs on the spring”.

Friedrich Herrmann

[1] *F. Herrmann:* Historical burdens on physics, Pure energy

1.15 The last secrets of nature

Subject:

„If physicists could answer these questions, we would finally begin to comprehend how matter functions at its most fundamental level. [1]

The technologies being developed today give us hope that by the time another 40 years roll around, we will have finally cracked the essential mystery of how matter, at its most fundamental level, is made.“ [1]

Similarly, but not so pompously formulated, it can be heard from other physicists: Sometime in the near future the puzzle will be solved, the code will be cracked.

When a journalist wants to explain it to the ordinary citizen, it may sound like this:

“With it [the LHC], so is the hope, eventually the ‘Theory of everything’ will be discovered – a formula that explains ‘whatever hold the world together in its inmost folds’.”

Deficiencies:

The “fundamental level”, the “essential mystery”, the “last secret”, and again and again the hackneyed quotation from Faust – all this is a manifestation of a desire: May the world be simple and clear-cut.

Apparently, scientists had always the hope, that their work will soon be completed. This can be seen from the names and designations they gave to the subjects and results of their work: the *indivisible* (átomos), the *elementary particle*, the *point* particle*,* the *first principles*, the *Theory of everything*. All these are concepts that cannot be surpassed, that express the end of something. The adjectives *elementary* and *fundamental* have no comparative or superlative; a size cannot be smaller than that of a point.

Such formulations come mostly from particle physicists. But this does not mean, that the underlaying convictions are not shared by solid state, plasma and other physicists, and also by the physical layman.

The idea of an end of the research efforts is by no means a new one. It had been promoted in every generation of physicists. The prospects for having understood the last and basic laws of nature were sometimes better, sometimes worse, but in general they did not really change. In particular at the end of the 19th century the future of science looked rather bright. With statistical thermodynamics and with Maxwell’s electrodynamics, which Maxwell himself conceived as a basically mechanical theory of the ether, the prospects seemed good for explaining in the near future all of the physical goings-on mechanically. In a lecture at the University of Munich in 1924 Max Planck explained how wrong was this point of view [2].

Today, i.e. another 90 years later, we have yet more demonstration material, that shows that such predictions do not become true. We know it because we have seen what came later. We know the future of the past. We do not know that of today, but if we want to learn something from the past it should not be only the insight that the conclusion that physics will soon be accomplished was premature, but also that such a conclusion would be frivolous also today, i.e. when we do not yet know the future. To express it more pointedly: We can learn from the past of science, that each solved problem generates at least one new unsolved problem. This extrapolation is not as risky as the expectation that the end will soon be reached.

Origin:

Probably several causes are acting together.

1. A simple explanation could be: Every human needs something that motivates him to do his work. When climbing a mountain one wants to be sure that there is a top of the mountain. Only reaching the top is the reward for the effort.

2. The expectation to catch on the ruse of nature may be the expression of a deep longing for sureness. Apparently good expert knowledge does not protect against naive expectations. The belief that there should be a definite and exhaustive physical explanation for everything may be compared with the belief in an almighty, whose almightiness does not require any further explanation. The question how god is working is considered illegitimate.

3. And finally: Marketing. Machines like the LHC or the RHIC are expensive. Projects of this size must be made attractive, digestible or acceptable to the general public. In other words: advertising has to be made. Otherwise, one might get the idea that the particle reactions in these colliders are not more interesting than a new chemical reaction was in the 19th century when it became clear that the huge diversity of substances can be traced back to a very small number of chemical elements. This was a great accomplishment, obtained however more cost-effectively.

The equipment needed by solid state physics, optoelectronics etc. are less expensive. The scientists working in these fields can promise us better data memories for our smart phones and similar improvements (that we have to pay for ourselves).

Not so in particle physics. It is particularly expensive, and solves problems generated by the particle physicists themselves. Finally we know where the mass is coming from, they tell us. The ordinary citizen may feel ashamed that he never had a problem with the concept of mass.

The great expense has to be justified, and that is why the transcendent is promised to the citizen. He does no longer really believe in the good Lord; so instead he gets the last, elementary, structureless particles or the Theory of Everything.

We should be indulgent with the science journalists when they exaggerate a little; they have to keep the readers in a good mood. However, perhaps it would not be bad if they also would feel responsible for the enlightenment of the public.

Disposal:

Particle physics is operating at the limits of the actual physical knowledge. It is expensive, but our affluent society can afford the huge colliders and telescopes that are required. But please: With another paradigm. Each solved problem is the origin of a new one. Is it so hard to endure this insight? Isn’t it the better motivation for doing physics, than the hope for an end of the scientific endeavor?

The enjoyment of having reached the summit would not last for a long time anyway.

This may remind us Sisyphos. But the comparison is not appropriate. Sisyphos had to roll the boulder up the same mountain again and again. The new problems of science on the contrary are new. Each attained altitude provides new views.

[1] *Sci. Am.* May 2015, S. 34f.

[2] *M. Planck:* Vom Relativen zum Absoluten, Gastvorlesung in der Universität München am 1. Dezember 1924, in „Wege zur Physikalischen Erkenntnis. Reden und Vorträge“, Band 1, S. Hirzel Verlag, Leipzig (1944), S. 142:

> When I began my physics studies and asked my venerable teacher Philipp v. Jolly for advice regarding the conditions and perspectives of my studies, he portrayed physics as a highly developed, and almost completely mature science, which now, after being crowned by the discovery of the principle of the conservation of energy would soon have assumed its definite form. It might be that in one or the other corner a dust particle or a small bubble remains which would have to be examined and classified, but the system as a whole should be secured, and theoretical physics approaches perceivably that state of achievement, that geometry has since centuries. This was fifty years ago the opinion of a physicist who was abreast with his times.

Friedrich Herrmann

1.16 The Hertzsprung-Russell diagram

Subject:

The Hertzsprung-Russell diagram is not missing in any upper secondary textbook. A point in this diagram represents a star. On the abscissa axis the spectral class or the surface temperature of the star is plotted, on the ordinate axis the luminosity.

Deficiencies:

1. The names of the axes

The variable of the ordinate axis is called luminosity. Actually, the luminosity is nothing but the energy flow.

The abscissa variable is often called a spectral type. One has the impression that this is not a variable, and if it is, not one that has a continuum of values. In fact, the spectra of stars are very complex. On the one hand, they are in good agreement with the spectrum of a Planckian radiator. On the other hand, they have both absorption and emission lines. From a need to bring order into the great variety of spectra, the spectral types have emerged, whose number and complexity has increased more and more over time. However, one can also characterize each spectrum by a single numerical value: the temperature of the blackbody radiation closest to the spectrum of the star.

For the learners, it would certainly be clearer to call the axes of the HR diagram “energy flow” and “temperature”.

2. The choice of the variables

But even more worrying is the choice of the variables: energy flow and temperature. Of course, the two variables correlate. However, what should actually be expressed? Normally, in physics, we are dealing with functional relationships, not correlations. Or spoken graphically: with lines, not with point clouds. We leave the correlative connections with their so-called scatter diagrams to sociologists, educators and economists.

In fact, one can also draw lines in the coordinate system of the HR diagram that describe a functional relationship: for the history of one single star, from its formation to its end, representing the flow of energy leaving it as a function of its surface temperature. What this looks like essentially depends on the mass of the star. So we have a set of functions with the mass as a parameter. In the HR diagram, on the other hand, the mass appears as a random variable.

But if we visualize the energy flow from a star as a function of its surface temperature, it immediately reveals what we are doing wrong. The message that we want to give to our students is quite another: every single star goes through an evolution. When we discuss stellar evolution in class, or in the lecture, we ask for functions of time, and if we want to look into the star, also from the position (in the form of the distance from the center of the star). For example, how does the temperature on the surface of the star (i.e., a fixed position r) depend on time, or how does the total energy flow outward depend on time.

Origin:

A classification of the stars had already been made in Hellenistic times. But the beginning of a physics of the stars, that is, of astrophysics, will be more in the 18th century. It was discovered that the stars did not rest, as previously thought and by the middle of the nineteenth century one was able to measure the distances of the nearest stars and thereby determine the absolute brightness of stars. Finally, the correlation between absolute brightness and the spectrum of the stars was discovered. The question about the source of energy was related to the idea that stars are going through an evolution, but at first this question could not be answered. The correlation expressed in the Hertzsprung-Russell diagram was one of the few observable phenomena of the time. From today’s point of view, one would rather classify this context in the category „raw data“, because what is actually physically interesting is the temporal evolution of a single star.

One of the reasons for the persistence of the HR diagram is certainly that it has its own name. If it has a name, it must be important, and Hertzsprung and Russell must have been important researchers. So no doubt: the HR diagram belongs to teaching objectives of a general education. However: What do the students learn about the many other researchers who made important contributions in this initial phase of astrophysics? Who discovered the movement of the stars? Who measured the first distance of a star? And finally, probably the most important question in this context: who had the idea that the energy source of the stars must be a nuclear reaction?

So again a typical example of the general topic of our column. On the basis of historical circumstances, a complicated, opaque scatter diagram was created; only a decade later it could be interpreted and could have been replaced by a more transparent representation. But it was not. The original description survived.

Disposal:

The evolution of stars is discussed for a typical sun-like star that ends up as a white dwarf; also one that ends up as a neutron star and one that ends up as a black hole. The Hertzsprung-Russell diagram is not needed.

Friedrich Herrmann

1.17 How to prepare it? How to detect it?

Subject:

The electric field is introduced via the force on a sample charge.

The electromagnetic wave is introduced via the open resonant circuit.

The wave function is explained by measuring the probability of finding the particle in a state with a sharp location.

The coherence of light is introduced via the properties of the light source.

Deficiencies:

The introduction of a new physical entity or "system" often starts with explaining how to realize or prepare or fabricate it, or how to prove its existence.

This has two disadvantages:

1. It often happens that students learn the process of preparation or that of verification. The properties of the object of consideration itself goes to the dogs.

   This is how they learn how to obtain field-line pictures and how a body exercises one force on another with the help of the field. Only the field itself remains vague and abstract. The question of the values of the physical standard quantities such as energy density, pressure, temperature, entropy of the field is hardly asked. It is as if we knew all the essentials about an electric field, as soon as we know what force is exerted on a test charge.

2. The explanation of the process of preparation or fabrication is usually more complicated than the description of the object in question. An example is the introduction of the electromagnetic wave via the Hertzian oscillator. It is extremely complicated and gives the impression that without Hertz's dipole you cannot understand the electromagnetic wave.

One encounters this kind of introduction in phenomena or systems that are considered complicated. However, one can ask oneself whether the impression of complexity is not caused by the indirect explanation.

If you were to explain to someone what air is, you certainly would not start talking about the formation of air in the course of Earth's history, and you would neither begin by proving the existence of the air by measuring its pressure.

Origin:

Due to historical coincidences, the system or phenomenon seemed complicated at the beginning. One believed that one could not talk about it in the same way as one usually talks about processes or physical systems. So there was a method that worked well in certain other contexts: the *operational definition*. Such a definition describes a procedure (an operation) with which to make or prove the phenomenon or the system.

Disposal:

Explain the electric and magnetic fields by talking about the properties of the field, especially about the energy density and the mechanical stresses in the fields, but also about its temperature and entropy.

Explain the electromagnetic wave by speaking the characteristics of a free-running wave. Start with the plane sine wave.

Explain coherence by speaking about the properties of coherent light. Speak about the light at that place where you want to characterize it (coherence generally changes from place to place).

Explain the wave function as long as it is not affected by a measurement, i.e. as long as it has been modified by a forced transition to another state.

Friedrich Herrmann

1.18 Female textbook authors

Subject:

"… it was decided to equip fast-moving locomotives with three-phase asynchronous motors. Power converters and thyristors (devices of power electronics) transform the single-phase alternating current picked up by the current collector first into direct current, then into three-phase current. The converters make it possible to feed energy released during braking as electrical energy of the correct frequency and phase back into the grid when the traction motors are switched as generators …"

Deficiencies:

I'd like to hope that every textbook author thinks that such phrases would never flow from his pen. And yet they come from a German school physics book. They are my demonstration quote, when I try in my lectures to show my students why it is not surprising that physics (together with chemistry) is one of the least popular school subjects. The sentence is not typical – thank goodness – but it is an extreme example of a style that is typical.

When teaching physics in the lower secondary school typically about 2000 technical terms are introduced. This corresponds to the basic vocabulary of a foreign language. I recommend to read the work of G. Merzyn on the relevant research [1]. According to Merzyn, in an average schoolbook one in six words is a technical term (in our above quote it is every third word). Half of all terms are used only once. Merzyn describes the predicament in moderate terms. I mean, an outcry would be appropriate! Regardless of its other qualities, a physics (or chemistry) book can not fulfill its purpose, as long as this flaw exists.

I admit, that there is a minority among learners who have the aptitude to recite this kind of babble – an observation that I occasionally make with students at the university. This can be quite advantageous: The more luscious the language, the less noticeable that one has understood nothing. Particularly conducive to the passing of an oral exam.

Origin:

A behavioral scientist would diagnose linguistic impersonation, as it is found in male primates.

Or can you, dear reader, imagine that the cited text was written by a woman? I do not know who wrote it, but among the 16 authors of the book there was not a single women.

Disposal:

Don't misunderstand me. I do not want you to embrace the learners with a youth language. The language should be simple and clear. A first step could be: Delete half of all technical terms from a textbook. That is not possible? Of course it is, see above: 50% of the terms are never used again. And once you've found fun in the jam, things are getting better, and you will see that you can reduce the remaining half once more. You will end up wondering how clear everything has become.

A practical suggestion for the authors: Give your text to read to your wife. She should mark every word whose meaning she does not know. (Of course, that could also be done by the editor of the publishing house.)

Another suggestion: Set an upper limit for the number of technical terms. This could easily be controlled when registering the textbook. Of course, it should be well below the number of vocabulary words in foreign language teaching.

Friedrich Herrmann

[1] G. Merzyn, Fachbestimmte Lernwege zur Förderung der Sprachkompetenz (3)

1.19 What is physics education research good for?

Subject:

The physics lessons at school and college follow in many details the historical development of the physical science.

Deficiencies:

The physics canon is largely unstructured and unnecessarily difficult, and it contains much that is superfluous.

Origin:

The physical science was developed – or evolved – essentially without a specific goal. Of course, there were motivations: the general search for knowledge and the practical pursuit of technical progress. But what the next discovery or invention would be could never be known in advance. And that's the way it is today: researching and searching in all directions. Much research doesn't give any result, but hardly anyone learns about such failures.

Nobody could foreshadow in Newton's time that 150 years later a field theory of electrical phenomena, electromagnetism, would arise. No one could foresee in the year 1800 that at the end of the same century statistics would become a fantastic tool in physics. Nobody would have believed at the end of that century that one was about to draw up a theory that called into question a basic principle of physics at that time, determinism.

But even on a much shorter time scale one can observe the unpredictability of physical results: In his famous work of 1905, Einstein explains that the ether is a superfluous concept. Ten years later he takes back this statement: „According to the general theory of relativity, a space without ether is unthinkable." [1]. Also in his work of 1905, Einstein, on the first three pages, elaborately explains that it is an important problem to synchronize clocks so that one can decide on the simultaneity of two distant events. Ten years later, with his theory of gravity, he shows that synchronizing clocks in a curved (i.e. a realistic) space is fundamentally impossible, and that in general one can not speak of the simultaneity of two events in different places.

One can compare the development of physics with the advance into a still unknown country, such as the American West in the early nineteenth century: One always advanced where it was most feasible. Only later did one find shorter ways and was able to built tunnels and bridges, so that a journey between the starting and ending point became much faster.

So physics has not started with the goal of reaching the state it is in today, but it always went the way that was just opening.

Now, this same zigzag path is pursued in an astonishingly manner in teaching physics – with grave consequences:

- It takes an unnecessary great amount of time.
- Structures and relationships that are only recognizable in retrospect are not presented in class.

One might think that it would be logical and reasonable to eliminate the shortcomings that such a zigzag course entails as soon as they appear. But that almost never happened. Why not?

The teaching of physics is anchored in a system that defies the smallest changes. This system includes teachers, university professors, textbooks, curricula, professional associations, and more. So it happened that physics became one of the most conservative school subjects.

In order to be better able to characterize the phenomenon, we want to ask the question of the time scales of various natural and social development processes.

For example: What is the typical duration of a war? (ten years). How long does a totalitarian regime survive? (30 years) How long does a clothing fashion last? (2 years). How long does a consumer habit, such as smoking, last?(100 years) How long does a weather situation last? (some days). How long does a religion hold? (1000 years). How long does it take to introduce a technical innovation? (5 years). And finally, how long does a teaching concept last?

The idea for this question and the estimated answers come from one of my colleagues in high energy physics. It was his way of characterizing the inertia of the teaching conventions. His answer to the last question was: 300 years.

Notice in particular the difference of the time scales for the introduction of a new teaching concept – some hundred years – and a technical innovation – some years or at most decades. One could have expected that both develop on the same time scale. In fact, the difference is huge.

Why can a technical development prevail so quickly, and a new teaching concept not?

The answer is straight-forward: In technology, strong competition ensures rapid development. The utility is measured in dollars, euros and yuan. Who does not progress, stays behind. The profit pays off in a near future, i.e. in a few years.

This type of feedback does not seem to exist in teaching. A textbook that is too innovative fails because it does not fit into the curriculum. The curricula can not be substantially modernized because teachers do not want to relearn and rethink. A university book that is too innovative does not have a chance with publishers because it does not sell.

And finally another obstacle: There is no corresponding research structure at the universities. Actually, one would think, it would be the task of the research in physics education, to question constantly the contents of the curriculum, to re-edit, to restructure, throw out superfluous subjects. However, there is a problem: such an activity is not appreciated by the researchers of other physical subject matters. And the researchers in physics education do not want to spoil it with those who use the longer lever. So one prefers to work either in the teaching-learning research. Thereby on does not harm any colleague from the physics department. Or one does something that you might call physical entertainment music. One examines and describes nice physical effects from sport and play and everyday life, and thus makes the promotional work for physics which is appreciated by our colleagues from the department of particle physics or nanoscience.

Disposal:

We have to experiment not only in but also with physics lessons. Only in this way can we find out which concepts meet the current problems.

This requires a competent and self-confident research in physics education, which not only focuses on what probably might think our colleagues who do the hard-core research.

Their task would be an examination of curricula and teaching programs of the universities, as well as a constant processing of the results of the current physical research.

Finally, an idea that does not seem to fit the mindset of us educators and researches in education. (I got the idea from a successful entrepreneur.) One tries to rate the lessons in monetary terms. Something like this: You develop a new lesson about an accepted content that gives the same results as an existing one. If less teaching time is needed for the new unit, it means an economic benefit.

Friedrich Herrmann

[1] Einstein, A.: Äther und Relativitätstheorie. Verlag von Julius Springer, Berlin 1920, S. 12

1.20 Transformations

Subject:

In physics, one often speaks of transformations. So energy is transformed from one form to another. But not only, as here, is one physical quantity transformed in the same physical quantity, but sometimes in another:

"… transform the charge into a voltage within the pixel using an amplifier circuit."

It can be even more complicated:

"It transforms the intensity and direction of the incident light into an electrical charge."

Finally, not only physical quantities are transformed or converted back and forth. Also, objects of the real world are transformed into physical quantities, such as when one says that light is transformed into energy, or as in this quote:

"The conversion of light into an electric charge is based on the internal photoelectric effect."

Deficiencies:

The quotes are not from the weekend edition of a local newspaper, but from the monthly Journal of the German Physical Society.

First a definition (from the *Wiktionary*). A "transformation" is "a marked change in appearance or character", to "convert" means "to transform or change (something) into another form, substance, state, or product".

Thus, transforming or converting means a process. Something was previously A and is later B. At the beginning it was not yet B and at the end it is no longer A. Like in the case of the wedding of Kana: before it was water and after the transformation it was wine.

A matter of course, trivial? Obviously not so trivial since in physics the term is not used correctly.

I do not want to comment the transforming energy here; it was already the subject of another article in his series. The same holds for the transformation of mass into energy.

I start with the charge, which is transformed into a voltage. Does the charge disappear in the pixel, and this causes a voltage? Probably not. And more striking is the disagreement in the quote where a direction is turned into a charge.

And even more it hurts to read that light is converted into energy, or that, as in our last quote, into electric charge. Not only physics, but above all logic speak is against this. How can light, i.e. an object of the real world, be converted into a physical quantity, i.e. a variable in the sense of mathematics?

One could reply, that this has to be accepted, since it is the physical colloquial language. That may be true, unfortunately.

You, dear reader, probably belonged to the 10 to 15% of students in school who are immune to bad physics lessons, and it was not difficult for you to cope with this somewhat inconsistent language. But you are only 10 to 15%.

One should not be surprised that in the minds of the students a conceptual chaos arises, and that they believe in the craziest conversions (based on my experience in exams, I can testify it): energy into momentum, momentum into angular momentum, energy into entropy and the like.

Origin:

As often, a sluttishness with the conceptual foundations of physics.

Of course, the precursor of all transformations is the conversion of energy, i.e. Work into heat, heat into work, heat in electric energy, electric in chemical energy etc.

The second cause is the confusion of statements about the real world and statements about the mathematical description of the world. Light can not transform into energy for logical reasons. Light **has** energy just as it has momentum, angular momentum and entropy. And the light has no electric charge, and it basically cannot transform into charge.

Disposal:

1. Pay attention to the language. Talking with a little more care does not mean that our statements become more complicated or difficult. On the contrary, they become clearer and easier to understand.

2. There are only a few situations in physics where you need the term transformation or conversion. Therefore my recommendation: dispose the word and the concept of transformation altogether.

Friedrich Herrmann

1.21 Deriving and understanding

Subject:

In the supervision of Bachelor and Master's theses I have made the observation that students like to "derive".

The tendency is also pronounced in the teaching at school: a newly introduced relationship between physical quantities must be derived or it must be deduced from an experiment.

Deficiencies:

This is not about science theory, but about something more modest: how to best understand a relation between physical quantities, or in short: to understand a formula?

One may have the impression that the main purpose of physics education is to prove the validity of a formula. If that is done, it seems, one has done one's duty; I mean the duty to make something understandable.

The "proof" can be made in two ways: 1. by deriving the formula, 2. by testing it in an experiment.

Now, based on my decades of experience in dealing with the students' problems, I can say that with the proof of a formula, the understanding has by no means been achieved. In many cases, despite the derivation, the learners do not yet have the slightest understanding of the considered formula. It even happens that, apart from the derivation, they understood neither the derived formula nor that from which it was derived. So they did not understand anything except the derivation.

In addition, in many cases, the derivation is harder to understand than the formula that is derived (as a smartphone or a car is easier to understand than its manufacturing process).

And if a derivation process is too difficult for the secondary school, unfortunately, the topic is completely omitted, although the result of the derivation could easily be understood. An example of this is the Fourier decomposition. A proof of the procedure is too complicated for the school, so the Fourier series are not treated in the classroom. But if you apply the method with the help of a simple computer app, the Fourier series can already be understood at an intermediate level.

Of course it is satisfying to derive the whole of electrodynamics from the Maxwell equations. The pursuit of axiomatics by the physicists, that is to say, of deriving all formulas from a few, which nature has given us without reason, is humanly intelligible. It is akin to the pursuit of „first principles", ultimate truths, most elementary particles, final equations, great unified theories.

Leaving the reins behind does not mean robbing the physics of their exactness. The formula that we simply write on the blackboard and whose statement we make plausible is a mathematical relationship, it is exact in the sense we want.

Origin:

One can assume various causes:

1. Computing replaces thinking. To repeat a calculation is less tiring than to work out an understanding of the physics behind a formula.

2. Even at the risk of arousing the wrath of my colleagues with the subject combination maths/physics: I fear that some of them are trimmed by their second subject, mathematics, to regard proving as the most important scientific activity.

3. Until not so long ago – I mean, as long as there were no computers – the analytical calculus was the most important tool for the exact description of physical issues. However, analysis as a tool of physics could soon suffer the same fate as geometry 300 years ago. At the time of Galileo, geometry was the only reliable means of accurately describing a physical phenomenon. („Who understands geometry can understand everything in this world" or „Nature speaks the language of mathematics: the letters of this language are triangles, circles and other mathematical figures.") This changed drastically after Newton introduced the differential calculus.

One might argue with Kant „… in every pure natural doctrine there is only so much of actual science as mathematics can be applied in it." This is certainly true, but mathematics is not simply the derivation of one from the other.

4. In physics teaching one usually aims at saying nothing without proving it. For Pohl, whom the elders among us still know from his classical textbooks on experimental physics, this quest was almost obsessive. The lecturer was not allowed to say anything in the lecture hall, which he had not demonstrated through an experiment. For me a question is whether the students in the lecture hall doubted with each new statement at its credibility. After all, the reputation of physics gives no reason of such a doubt, unlike a number of other subject matters in which one school of thought, fashion or ideology follows the other, and where one can hardly come up with derivations or experiments.

Disposal:

The most important thing to do when introducing a new formula: discuss the formula itself, so that the students at the end have the feeling that they could have written down the formula themselves.

Here is a simple example from the school: the formula

$$E_{\text{kin}} = \frac{m}{2} v^2$$

for the kinetic energy. The derivation from another, familiar equation, such as

$$\Delta E = F \cdot \Delta s$$

is complicated because one has to integrate, and the integral calculus may not yet be available. In addition, the integration corresponds to a physical process that does not matter in the end.

In fact, one can obtain the equation, or at least its essential part, without calculation, but only with an educated guess. First of all, you realize that the energy you are looking for depends on the mass and on the velocity, and on nothing else. Then one easily convinces oneself that the energy must be proportional to the mass, because the energy is a substance-like (extensive) quantity, and therefore on the right side of the equation there must also be a substance-like quantity in the first power. (Two bodies of the same mass must have twice as much kinetic energy at the same velocity as a single one.) Finally, the dependence on v. The energy is certainly independent of the direction of the movement, or in one dimension, of the sign of the velocity. The simplest function that can be used for this purpose is v^2. Even for the factor 1/2, there is an argument. In fact, if other energy formulas have been discussed before, this factor has already been encountered: in the energy stored in a spring when tensioned, in the energy in the capacitor or in the coil.

Yet another recommendation for the derivation at the university, where the number of calculation steps can be significantly greater: Try to interpret each intermediate result, because every intermediate result makes a physical statement.

Finally, an alternative to derivation: Modeling systems. In my opinion, they are not used enough in physics lessons. Dealing with them is easy to learn. They provide a good understanding and lead to a logical penetration of physical processes, by freeing us from the effort of the calculation.

Friedrich Herrmann

1.22 The I, the observer and the good Lord

Subject:

In physics, the “observer” plays an important role, unlike in chemistry or biology. In conversations about physical phenomena the observer often becomes the “I” or the “me”. “I see the yardstick shortened”, “for me the half-life of the muon is …”, “when driving around a sharp turn I am pushed aside”.

In quantum physics, the role of the observer is even more important: He (yes, the observer seems always to be male) is the one who likes to make measurements, and thereby somehow disturbs the system on which he is measuring.

Deficiencies:

In physics we use two perspectives of the world. Let us illustrate them with a simple example, namely the idea we make of the Earth.

The Earth from the normal all-day perspective:

We only see a small part, limited by the horizon. In addition, objects that are far away appear small; what is closer, seems bigger. I can see the bell tower of the neighboring village from my house. It appears to me at an angle of 1°. The church tower of my own village I see under 20°.

If one takes this viewpoint in physics, one speaks of “observation”; you yourself are the observer.

The Earth from the perspective of the knower:

In our mind, an idea arises that is quite different: the Earth is a sphere; the two church towers are the same height; both Europeans and Australians are standing with their feet on the ground ... One could also say that is how God sees the Earth, and we try to do it the same. The good Lord does not need a particular observational location; for him there is no horizon, not even a specific instant of observation. And when he imagines a quantum mechanical system, he does not bother it with a measurement.

Which of the two perspectives is that of physics? One might think that it is the concern of physics to see the world as the good Lord: not limited by horizons; the symmetry is not broken by the arbitrariness of coordinate systems and other reference frames; the hydrogen atom is not disturbed by a measurement.

But that’s not the way physics is, and that’s not how it should be. Because physicists also want to know what a human being sees – a person whom they like to call observer; a “me” so. Physicists also have to make statements about how to test their claims. The yardstick is what it is, but physics should also tell us how it appears. It looks smaller from a distance, and it also looks smaller from close up if it moves quickly.

However, one consequence of this is that the description of the world becomes more complicated than it would be without this requirement. Since all of us, and also the physicists, are human and not God, the description of the effects that have to do with our observation is essential. So what then is wrong?

I believe that the point of view of the observer, the experimenter, the „I“ plays too great a role in our physical discourse and especially in teaching.

A considerable part of the difficulties that everyone has with the special theory of relativity comes only from answering the question of how the length of an object appears to someone, or the length of a time interval between two events. A significant amount of class time is spent describing the artifacts that result from choosing and changing the frames of reference. By the way, a fatal consequence in this particular case is that the learner is unaware that in the same context, i. e. in the theory of relativity, real changes in length occur: e.g. the distance between the mirrors of a gravitational wave detector.

Things are similar in quantum physics. The hydrogen atom could be so simple. But one also wants to say what an observer „sees“ and measures. And of course, the observer is assumed to want to know exactly where an electron is at any moment. So the observer will make a measurement of the position of the electron, thereby destroying the beautifully simple state of the undisturbed atom.

Yet another example: Matter crashes into a black hole. We, the outside people and observers, „see“ that the matter falling towards the Schwarzschild horizon is getting slower and slower, never reaching the horizon. An imaginary observer who is falling together with the matter in free fall flies through the horizon, without noticing anything of it. How does that fit together? It’s not much different from the church towers. If you want to know what the world is like, do not ask what the observers see, but ask for the physical description of the object itself – it tells you everything. Of course, it also tells us what the various observers see and experience, but these are just details. This latter information serves less to understand the world than the physical craft.

Origin:

Possibly from the traditional positivist attitude of science: one only accepts what can be measured and verified. That’s a reasonable attitude. To a certain extent a hygiene behavior that is a prerequisite for physics to be able to make more authoritative statements than certain other fields of science. But it is certainly just as reasonable to assume that things, the objects of our consideration, the physical systems, exist even if we do not observe and measure them.

By the way, it was worse in the past. The older ones will remember that when we were students, we not only had to know what the observer measures, but we also had to be able to explain the operation principle of the meter: the galvanometer, the dynamometer, and the Geiger counter.

Disposal:

Deal sparingly with the term observer. Prefer descriptions that are independent of the observer. Of course, the “me” can make a thing clear, especially in the oral presentation. But actually it should not occur in our arguments (unless, one discusses the momentum exchange, for example between “me” and “you”).

Especially in the special theory of relativity let the Lorentz transformations first aside, and in quantum physics, the measurement.

Friedrich Herrmann

1.23 The one and the other electron

Subject:

The Hamiltonian of a many-electron atom contains the coordinates of the individual electrons: r_1 belongs to electron 1, r_2 belongs to electron 2, etc. The Pauli principle requires that the wave function of such many-electron systems be antisymmetric: by exchanging two particle coordinates, it changes its sign. In the case of a two-electron system, assuming that the two-particle wave function can be written as the product of two one-particle wave functions, one has

$$\psi(1,2) = \frac{1}{\sqrt{2}}[\psi_a(1)\psi_b(2) - \psi_a(2)\psi_b(1)]$$

The 1 and the 2 stand for the coordinates of the two electrons, a and b stand for the states of the two electrons.

Deficiencies:

In this context it is said that one has to deal with an electron in state a and one in state b (e. g. with an s and a p electron).

That sounds familiar; this or something similar reads in the chapter on many-electron atoms in any physics or chemistry textbook. But there is an inconsistency, and the text elegantly ignores it.

The problem always arises when one speaks of one single electron in a many-electron system, that is, electron 1, electron 2, an 2s electron, a 3p electron, or a 4f electron.

The language one uses is that normally used when one speaks of a well-defined individual entity. And anyone who reads the corresponding statements about the electron interprets them as follows: electron 1 is one particle, and electron 2 is the other one – even though one can not say exactly where the particles are. You learn then, that they are indistinguishable, which is certainly not easy to understand. But once you have accepted it, you have two electrons 1 and 2 – not distinguishable, but still one is electron 1 and the other electron 2.

Now the problem: Shortly thereafter several electrons are mentioned. This time, they are called a and b, or 3d, 4f, etc. However, it may not have been noticed that the electrons 1 and 2 are not the same as electrons 3d and 4f. Keep in mind that in both cases we speak of the same atom in the same state.

The mathematics that lie in between, certainly shows what has happened, and what is the relationship between 1 and 2 on the one hand, and a and b on the other. In the language used, however, this relationship is not reflected.

In fact, you are in a similar situation as with two coupled pendulums. Let's call them pendulum 1 and pendulum 2. Then we describe the solution with two normal modes a and b. Nobody would come up with the idea here of calling the two normal modes pendulum a and pendulum b.

Origin:

The language is that of the Bohr model: an electron is a small particle that "circles" around the nucleus. Neither the so-called indistinguishability of the particles nor the Pauli principle fits in with this idea. But now the language was there, the idea of the electron as a small individual body has settled in our minds. And we had to accept incongruities that result only from the models transported by the language.

Incidentally, the idea of the electron as an individual is not always bad: depending on the state of the particle, it can become asymptotically as good as you like.

Disposal:

The disposal is difficult. Either one would have to use a language that depicts only the mathematics, and which is not based on a model. Or a language based on a model that fits a bit better than the Bohr model. It might be the model of the electron liquid, electron matter or electronium.

According to such a model, an electron is not an individual but a specific portion of a "substance": a portion of mass m_e and electric charge e.

Of course, also this model is not always applicable: it does not represent the interference. But its advantage is that it contains the fact that an electron has a certain electric charge and a certain mass, and it does not require an unintelligible, even unreasonable explanation like the one that says that the electrons are indistinguishable.

Friedrich Herrmann

1.24 Mass and matter

Subject:

“A black hole differs dramatically from a star of any other kind. Other stars contain both matter and mass. In contrast, the black hole is disembodied mass, mass without matter.”

“At the center of a black hole is the point of *crunch*. There the matter that once composed the star is crushed out of existence. In that crunch matter disappears, with all its particles, pressures, and properties. Pure matter-free mass remains.”

“Part of the matter is transformed into energy.”

“Transport of matter, charge and energy ….”

Deficiencies:

I don’t want to make a contribution to the centuries-old philosophical discussion about the term matter. I am also not interested in the delimitation expressed in terms such as “mind and matter”, “matter and field”, “light and matter”. My subject is a simpler question.

Matter is a part of the real world, something that was not invented by man, which also existed if no one had invented a name for it.

Mass, on the other hand, is a physical quantity, a variable in the sense of mathematics, introduced by man to describe certain properties of objects, namely their heaviness and their inertia.

Anyone who agrees with this simple statement will realize that the above quotes are not just awkward. They are logically not consistent; they have no meaning. Pure mass is meaningless. Mass is a measure. The measure without an object or entity to which it refers is meaningless.

A sack of potatoes has a certain weight. The weight without the potatoes is pointless.

Such a thing only exists in the fairytale world: The grin of the cat without the cat in Alice in Wonderland [1], – but that was possible only for a brief moment, and Alice is surprised accordingly.

However, one will not be too surprised about the quoted sentences if one has already noticed that it is often said that light *is* energy, or that photons are energy quanta.

With my students I like to do the following game. “We imagine an electron; right in front of us, in peace”. (No one seems to have a problem with the fact that this contradicts quantum physics). “Now we want to take away the electric charge of the electron, only in thought. Can you imagine that?” “Yes, we can; then we have an uncharged electron.”(One should actually give it a new name, but we’ll leave it by the name electron, because that tells us how the new entity came into existence.) I go on asking, “The electron has a spin, one can roughly imagine that it rotates. We now want to take away the spin. Can you imagine that?” “Yes, we can; we obtain an electron without charge that does not rotate.“ I skip the other extensive quantities characterizing the electron, such as the leptonic charge, and come straight to the mass:“ “We now want to take away the mass from the electron. Can you imagine that?” “No, that will not work. Then there is nothing left of the electron.” “But we could perhaps give it back its angular momentum and take away the mass. Is that possible?” “No, there is nothing left to rotate.” Etc., etc. *

What we took away in this game were always extensive quantities. More precisely: we set their values to zero.

I would not be surprised if the reader would shudder while reading these lines, but still: Don’t they express something, which has a certain plausibility?

The amazing thing is how much you can take away without our minds resisting it. But with the mass it seems to have an end: One has the feeling that one not only reduces the value of a quantity to zero, but one removes the considered object itself, the proper electron, or perhaps its soul. In the case of the particles, such as the electron, the soul always appears to be in its mass, or more precisely, in its rest mass.

Origin:

There is certainly some metaphysics. I can only guess what’s going on in the minds of some of my colleagues.

Perhaps the following: A body, a particle, a piece of matter is an individual. It is more than the ensemble of the values of its physical quantities. The idea may be that the amount of this metaphysical stuff is measured by the mass. When the mass is changed, the body is no longer the same as before. It remains the same when its momentum changes, or its angular momentum or entropy, and perhaps its charge, but not when its mass changes. The mass measures the amount of matter, and matter is something that goes beyond the physical quantities.

This might also be the reason why the particle physicists use the word mass for the rest mass, i.e. that part of the mass which does not change with a change of the state of motion.

Disposal:

Carefully distinguish between object and physical Quantity, or between “object and measure”. Try not to impute something to the mass, which is more than a measure of a property.

Friedrich Herrmann

* If one takes a macroscopic body instead of the electron, one seems to have fewer problems at this point. Everyone knows the massless spring of a spring oscillator or the massless thread of a pendulum. These ideas are used as carelessly as the resistance-free electrical conductor or the frictionless rolling car. Apparently, one behaves much more enlightened at this point than when trying to imagine a massless electron. Incidentally, our game is interesting also if you do it with a photon.

[1] Alice’s Adventures in Wonderland, by Lewis Carroll:

“All right,” said the Cat; and this time it vanished quite slowly, beginning with the end of the tail, and ending with the grin, which remained some time after the rest of it had gone.
“Well ! I’ve often seen a cat without a grin,” thought Alice ; “but a grin without a cat! It’s the most curious thing I ever saw in all my life!”

1.25 Broadband internet access

Subject:

Gobbledygook proliferation (using spectroscopy as an example)

Deficiencies:

The *Telekom* company informs us: “LTE can generate a bandwidth of up to 150 Mbit/s in simple operation. In cities even up to 300 Mbit/s are possible.”

Scientific American writes: “…a laser emits a narrow band of frequencies at best.... But having a light source that combines the properties of a laser with the broad bandwidth of an incandescent bulb opens up a whole new realm of possibilities.”

And so on, and so on... We come across the terms again and again: broadband Internet, bandwidth, line spectrum, spectral bands, band-pass filter,…

Where do these names come from? They refer to what we see when we decompose light with the help of a slit and a prism or diffraction grating. Depending on the type of light source, one sees on a screen optical images of the slit, which are shifted against each other according to the wavelengths. The whole image thus consists of more or less wide lines or bands. One is interested in the energy flux density per frequency interval. Therefore, these images are analyzed and the result is displayed over the frequency as a graph. This is at least what it is today.

The terms line and band therefore refer to the raw data, an artifact resulting from a particular technical arrangement. When one speaks of a band or a bandwidth, one means a frequency interval, or as in our *Telekom* quote a data transmission rate.

One might object: But that’s the way it is – this way of speaking has become established, everyone understands it, that’s how language works.

There’s nothing wrong with that at first. However, one might ask oneself: If the words “line” and “band” are not intended to do more than characterize an interval of a physical quantity, why this particular wording? Why does a frequency interval need a name of its own? In the same way, one could call an electrical potential difference a “voltage sector”, because the pointer of a voltmeter covers a sector. Or a time interval might be called an “time angle”, because on the dial of a watch it corresponds to an angle.

But once again, is it worth talking about it? If it were only the lines and the bands, the subject would actually be uninteresting. The problem is not the individual case of the frequency band, but the fact that physics lessons (as well as chemistry lessons) are overloaded with clumsy, thoughtless, unnecessary special formulations.

So the question is: Why do we use ever new words for something we can easily say with the old ones?

Origin:

In the early days of spectroscopy (to which we owe almost all the findings that eventually led to quantum mechanics), the observation result was a “spectrum”, i.e. the image of the diffraction slit; it was not a graph of the spectral energy flux density over frequency. The corresponding modes of speech have become established and widespread.

Disposal:

As far as our example is concerned, it is not difficult to express oneself more appropriately: frequency interval instead of band, or in the case of the Internet access: Internet access with a high data rate. Or even closer to the colloquial language: fast Internet access,...

My real concern, however, is something else. Namely, a reference to France.

There is a tradition of language cultivation in France. In Germany, on the other hand, the motto seems to be: language is what is spoken, free language for free citizens.

We would need an institution that thinks about what language is appropriate in a newly emerging scientific or technical context. The aim should be to make suggestions for an unpretentious, clear language, and to throw out superfluous and unclear terms, just as one removes weeds in the garden.

By the way: broadband access to the Internet is called *Accès à haut débit* in French. It couldn’t be said more clearly.

Friedrich Herrmann

1.26 Keep it simple…

Subject:

1. A carriage is coasting to a stop.

Physical description: The earth or the ground or the road or the air exerts a force on the carriage. This causes a negative acceleration of the carriage.

2. A (positively) charged conducting sphere is discharging.

Physical description: Electrons flow from the earth via the conductive connection onto the sphere so that the sphere becomes neutral.

3. A cup of hot coffee is cooling down.

Physical description: The internal energy of the coffee decreases as it releases energy in the form of heat to the environment. As a result, the enthalpy of the environment increases.

Deficiencies:

The three processes are largely analogous to each other. In each of the three cases one has a flow of an extensive quantity and a gradient of the corresponding intensive quantity. In each of the three cases, the extensive quantity is no longer noticeable after the process has come to an end.

It is, so to speak, three times the same play, performed with different actors: In the first case, momentum and velocity, in the second, electric charge and electric potential, and in the third, entropy and temperature.

This is how simple the good Lord (or whoever) made the world, but people have not yet realized it.

Origin:

The processes were first described in very different epochs, over a period of up to 100 years, under very different conditions, by different researchers at different places. When finally the similarities could have been seen, or were even seen, it was already too late. Nothing changes in an explanatory pattern that is firmly anchored in textbooks, curricula and in people's minds – not necessarily because one is not able to change something, but probably because one does not want to do so. If someone does dare to point out the possibility of a simplification, the council of the relevant "religious community" will decide on measures by which such behavior will be sanctioned.

Disposal:

1. Momentum flows by itself from the body with the higher to the body with the lower velocity.

2. Electrical charge flows by itself from the body with the higher to the body with the lower electrical potential.

3. Entropy flows by itself from the body with the higher temperature to the body with the lower temperature

In all three cases entropy is generated, and energy is needed to transport it away.

In the end, one does not notice anything of the momentum, charge or entropy, because the systems that have absorbed these quantities is very large. So they are greatly diluted.

Friedrich Herrmann

1.27 Lack of terms

Subject:

Our textbooks and also our lessons undoubtedly contain too many technical terms [1]. But there are also important concepts in physics for which we have no names. Thus, technical terms are missing.

Deficiencies :

We discuss three phenomena where the lack of a designation is of similar nature.

1. Fields

The word is used in physics with different meanings: firstly as a name for the distribution of a local physical quantity in space. For example, one speaks of a temperature field or a flow field. We are not concerned with this meaning here, but with the other one, according to which field is the name of a physical system, i.e. of an entity which exists in nature, no matter whether we describe it mathematically or not. Examples are the electric field, the magnetic field and the gravitational field.

In this sense, a field is an extended entity. If one refers to a certain field, for example that of a magnetic dipole, one can say where it is located – not at a point, but in a region of space. The fact that it doesn't have a sharp boundary need not bother us. We also speak of the earth's atmosphere and everybody knows what is meant: Somewhere its density is so low that one can say: it reaches to about here.

Now the problem: If one has a clear idea about the field, it happens that one wants to address its local properties, for example its energy density or its mechanical stresses. So one will say: the energy density at a certain location within the field is.... This is clear, but it is messy. What is wrong with it? One would probably not say, the temperature of the atmosphere at a certain point, but the temperature of the air at this point. One would like to refer to the materiel or the substance, and not the extended entity.

In the classroom we have clearly felt this lack, and decided to introduce a name for it: Field stuff. (I admit that it is not very original).

It is interesting that in another context, namely in thermodynamics, one has a local designation for the electromagnetic field: radiation. The whole entity can then be called radiation field.

2. Space

If it is our learning objective to make clear that space has local properties, it would help the understanding to have some kind of substance name. Of course, the name ether comes to mind here. One could then say: the ether here is differently curved than the ether there.

3. The electron

We first recall that in physics one of two "extreme models" is used in many contexts: on the one hand the particle model, on the other the continuum or substance model.

Particle model means: One imagines certain physical objects, called particles, as point-like. Thus, in the model, they have no extension.

In the substance model, all objects have an extension, and the "substance" is continuously distributed in space and has (local) properties at each point.

Both models make statements which in principle cannot be confirmed or disproved. They are metaphysical statements. We use models, i.e. ideas about the world which allow to draw conclusions that are correct in the relevant context – provided that one has chosen the appropriate model for the description and one has not overused it.

In fact, entities which are usually called particles can be described with both models, e.g. electrons. Depending on the phenomenon to be described, one or the other is the more suitable. If one wants to describe the electron shell of an atom or the electron system of a solid, the substance model is more suitable than the particle model. To give just one example of the advantage: Instead of introducing the "Bohr postulate", according to which an electron circling around the nucleus does not emit an electromagnetic wave, i.e. one suspends electromagnetism for a moment, one describes the situation better by a ring current of a substance distributed around the nucleus, which, however, has no name. One can easily specify the values of local quantities: Mass density, charge density, the corresponding current densities, even a velocity. Thus one also has an explanation of the angular momentum and the magnetic moment of the electron distributed in space in a state with m unequal to 0. The handling of this model is greatly facilitated if one introduces a name for the "substance" which has these properties.

Origin:

It is noteworthy that a deficiency of a similar nature occurs in three completely different contexts. The origin in the three cases is different.

In the first case: A field is often reduced to its mathematical description by a field strength. According to this, a field is a spatial area in which forces can act, if there is a test body. Such a treatment of the field does not suggest the question about the values of other local physical quantities. The question of a name for a local entity rather does not arise.

We will not go in detail into the long and complicated history of the concept of space and the ether here.

Concerning the electron and other particles: Their history was always connected with the search for the last building elements of matter, and apparently these could only be imagined as point-like. If they had an extension, they would have to have an inner structure and could not be the last, indivisible, elementary - this may be the somewhat naive conclusion. Perhaps the well-known infatuation of physicists with point mechanics also plays a role.

Disposal:

We would like to encourage teachers to use substance names, especially for the electric and the magnetic field (field substance), and also for the electron. The old name *Madelung liquid* is a bit bulky. We use the designation *electronium* in our lessons.

Apart from the bad reputation the ether has got because of its misuse, there are two other things to consider. First: the name we are looking for is only usable for the three-dimensional space. It is not suitable for space-time. And secondly: A substance is always imbedded in space, or more simply: in a container. But the space we are talking about has the property that it is container and content in one. For such a structure we have no model and therefore there is probably no suggestive name. But it helps to discuss exactly this point in class [2].

[1] G. Merzyn, Fachbestimmte Lernwege zur Förderung der Sprach-kompetenz (3)

https://www.schulentwicklung.nrw.de/cms/upload/sprachsensibler_FU/Fachbestimmte_Lernwege_zur_Foerderung_der_Sprachkompetenz_Naturwissenschaften_Mercyn.pdf

[2] *The Karlsruhe Physics Course for the secondary school A-level: Mechanics*; 9.1 Space – more than an empty recipient

Friedrich Herrmann

1.28 Shelf warmers

Subject:

The first law, the third law, the zeroth law, Newton's first law, Newton's third law, Lenz's rule, and many others.

Deficiencies:

There are laws that you learn as a physics student, and which are presented as important, but which you do not need. These include:

- The first law which is not needed if one knows that energy is a conserved quantity and if one has a sound understanding of entropy.
- The zeroth law which is not needed if one has a sound understanding of entropy.
- The third law which is not needed if one knows that energy is a conserved quantity and if one has a sound understanding of entropy.
- Newton's first law is not needed if one knows Newton's second law.
- Newton's third law is not needed if one has understood the concept of force (preferably if one has understood that forces are momentum currents).

Equally superfluous are some physical quantities, derivations, descriptions of experiments and other subjects.

The educational canon (of university and school) contains topics that are superfluous and make physics seem more complicated than it actually is. Their treatment requires time that we could urgently use for other topics. They are like shelf warmers: shelf warmers need space and care, they have to be cleaned from dust, and they do not bring profit for the store, because nobody wants to buy them.

Origin:

First the original shelf warmers: business with them was good at one time, but the demands of buyers have changed, clothing fashions have changed, or the device has become obsolete because a more modern version has come onto the market. And so it is with physical shelf warmers. Once they were useful, but they are no longer.

Disposal:

Shelf warmer items are taken out of the assortment. And we should do the same with many of our former physical gems. It is easy to get over the loss. After a short period of getting used to it, you feel really liberated.

Friedrich Herrmann

1.29 Definitions

Subject:

To talk about a physical process or context, there must be clarity about the meaning of the concepts used.

It may be that the meaning of a term is known. If not, we have to explain it, for example with a definition, like the one at the beginning of the corresponding entry in an encyclopedia.

A definition should clearly state what it includes and what it excludes. As a teacher, we like to write the definition of a new term on the board as a mnemonic.

Deficiencies:

It often happens that such a definition is not possible. If one nevertheless tries to define, this has undesirable consequences. We usually find ourselves in the following dilemma: We try a formulation, but find that it includes situations, systems or processes that we do not want to be included. So we narrow it down. The result: Now it no longer contains alien elements, but it excludes things that it should not exclude.

We consider as an example the concept called in physics oscillation. So: What is an oscillation?

Here are some examples from high school physics textbooks:

1. “Processes of motion in which the direction is repeatedly reversed and which seem to repeat after a certain time are called oscillations.”

2. “A mechanical oscillation is a time-periodic movement of a body around an equilibrium position…. An oscillation is a time-periodic change of physical quantities.”

3. “We encounter periodic processes in many different ways. Whether a swing, a guitar string, the voltage at the socket, or the blood pressure in our veins, everywhere a physical quantity changes in a certain rhythm. When a physical quantity ‘oscillates’ back to a certain value and beyond again and again, this is called an oscillation. The oscillating objects are called oscillators.”

Or from a university textbook:

4. “Oscillations can result when a system is slightly deflected from a stable equilibrium position. What is remarkable about oscillations is that their motion is periodic, that is, repetitive. Many oscillation phenomena are familiar to us: the up and down of small boats, the back and forth of clock pendulums, and the oscillation of strings and reeds in musical instruments. In addition, there are examples of oscillations that are not so familiar to us: the oscillations of air molecules in sound waves and the oscillations of electric currents in radio and television sets.”

Or from Wikipedia:

5. “Oscillations (Latin oscillare ‘to rock’) are repeated temporal fluctuations of state variables of a system. An oscillation is a deviation from an average value. Oscillations can occur in all feedback systems. Examples of oscillations can be found in mechanics, electrical engineering, biology, economics and many other fields.”

Let’s review the quotes:

Re 1: The definition is too broad. One would not want to call every motion sequence that repeats itself after a certain time an oscillation, e.g. the motion of a streetcar between its two terminus stations.

Re 2. The requirement of periodicity in time excludes damped oscillations.

Re 3: Probably the author has noticed that a definition is not simple and has drawn a consequence. He simply declares everything to be an oscillation which is periodic. However, he does not meet what in physics is understood by an oscillation: Periodicity alone does not make an oscillation. And the damped oscillation would be excluded again. Afterwards the term oscillator is explained: an oscillator is a vibrating object. But who is the oscillating object in the examples given? For example the voltage at the socket. Is the voltage the object, or the socket? Is the blood in the veins the oscillator?

Re 4: Also by this definition damped oscillations are excluded. And by the way: Air molecules move with about 500 m/s. Between two collisions they move with constant speed. However, the speed that belongs to the oscillatory motion is only about 0.5 mm/s. To call this process “oscillation of molecules” certainly does not meet what is meant by the term, neither in physics, nor in colloquial language.

Re 5: Again, a very broad definition was used so that it includes periodic processes that would hardly be called oscillations in physics.

None of the definitions also makes it clear whether what is called a forced oscillation in physics is covered.

When defining something, one should imagine one is doing it for someone who does not yet know the concept. This is not the case with any of the quotes.

We in no way wanted to use these quotes to demonstrate the authors’ incompetence. Rather, we wanted to show:

- how difficult it is to formulate a definition.
- that obviously the definition is not needed, because in spite of the failed attempts to define it, everybody knows what is meant by an oscillation in physics.

Origin:

Here we are not concerned with the origin of the concept definition, but with that of the desire to define. It probably originates from our tendency to seek certainty. Our need for one-bit statements is probably anchored in our genes. Humans (and animals) often have to make “one-bit decisions”, such as: flee or stay.

That’s why attributes that are continuously changing are often projected onto yes-no-attributes. “Good or evil” (you can see it in the current political discussions) or “What is life?” One just wants a simple truth.

Disposal:

The fact that the edges are blurred is normal for the terms of the everyday language. As soon as one wants to make the edges sharp, things become ugly, long, incomprehensible, see, for example, the legalese.

In everyday life, however, communication works well without sharp definitions, or even because of their vagueness.

A child learns the meaning of concepts not by definitions, but by examples and by analogies.

And finally, we are currently witnessing that computing, which until recently was based on clear yes-no decisions, is developing a new power in the form of neural networks and artificial intelligence, i.e. methods that operate without clear definitions. The computer is trained, it learns like a child: without definitions, but by examples.

What does this imply for defining in the classroom?

Be cautious about defining. Instead of definitions, examples are usually the better choice.

And in the special case of oscillations: Address phenomena that are intermediate between oscillation and non-oscillation: this may cost a little more time, but will have been worth it in the end.

Friedrich Herrmann

1.30 The independent variable

Subject:

Often entropy (as well as enthalpy) is plotted against the temperature, like in Fig. 1.

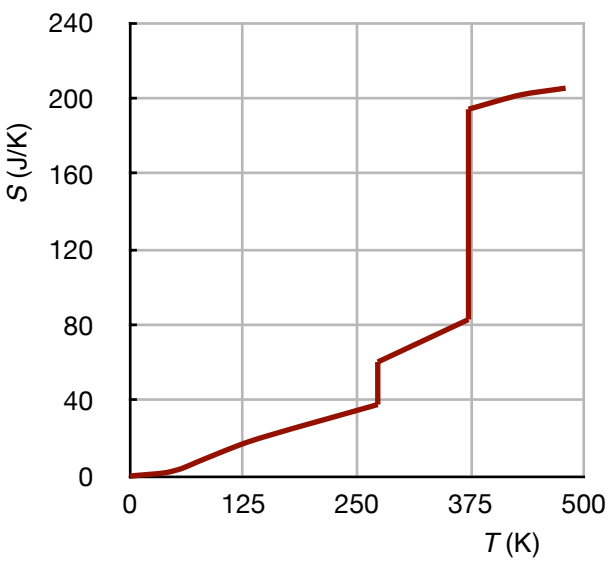


Fig. 1. Entropy of 1 mol water as a function of temperature.

At the temperature of a phase transition, the entropy is making a jump.

Often momentum is plotted against the velocity as in Fig. 2.

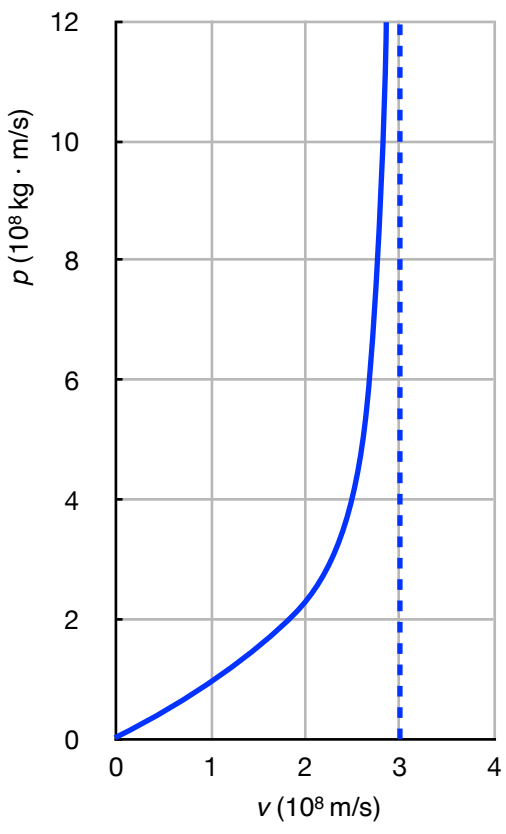


Fig. 2. Momentum of a body of mass 1 kg as a function of velocity.

At the limiting velocity *c*, the curve has a pole. There is a singularity.

Deficiencies:

Both figures suggest something that doesn't quite hit the mark. Perhaps the old *Natura non facit saltus* comes to mind, but the matter is even simpler. It is enough to reverse the assignment of the two variables to the abscissa and ordinate axes respectively, figures 3 and 4. What was the independent variable before becomes the dependent one, and vice versa.

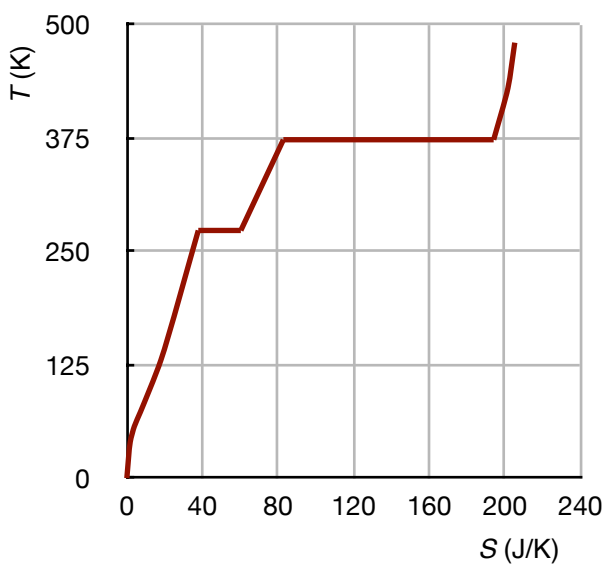


Fig. 3. Temperature of 1 mol of water as a function of entropy.

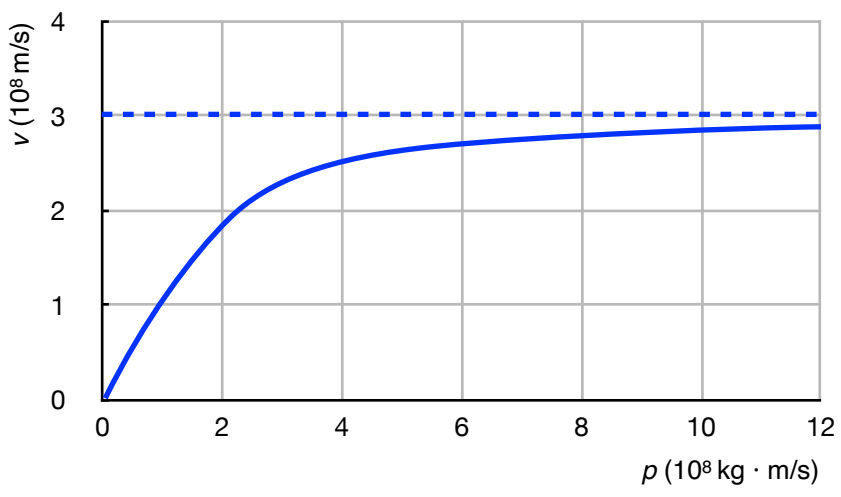


Fig. 4. Velocity of a body of mass 1 kg as a function of momentum.

Both the jumps and the pole, or singularity, have disappeared.

But isn't this just a harmless trick without any consequence? Not at all. The diagrams do not only inform us about the relationship between *S* and *T*, or between **p** and **v**. They also tell us which is the "independent" variable, and by independent is meant: It is up to me to choose its value. Therefore, let us look at the two cases again.

Fig. 1 tells, or suggests: change the temperature, then the entropy content of the system under consideration changes as shown by the curve. In particular, it tells us that the entropy increases abruptly at 273 K.

Or Fig. 2: change the velocity, and you will notice that as you approach 300 000 km/s, the momentum increases more and more and takes on gigantic values.

The problem is that it is not possible to simply increase the temperature from 273 K to 274 K. It simply does not work. Apparently the independent variable is not as independent as we thought. How did we manage to increase the temperature in the first place? By adding entropy! What we can do is add or remove entropy. The body then reacts to it in the way it likes to do.

The same is true for momentum and velocity. The quantity that we can easily handle is the momentum. We press the accelerator pedal so that the engine pumps momentum from the earth into the car. At the speedometer we see the effect. Or we charge a particle with momentum: the particle initially speeds up, and as the supply of momentum continues, the increase in speed becomes smaller and smaller. In a particle "accelerator", most of the momentum supply takes place after the particles have already reached the limiting velocity, so that their velocity (almost) does not change any more, there is no acceleration.

Once again: when we heat, we add entropy; how the body reacts is up to the body. It can get warmer, but it does not have to. When accelerating, we add momentum. How the body reacts is its own business. It can become faster, but it does not have to.

Origin:

One chooses as independent variable the quantity for which one believes to have a better understanding. In mechanics, this is, of course, the velocity, especially after three weeks of teaching kinematics, and by degrading momentum to an auxiliary quantity for the description of collision experiments.

In thermodynamics, temperature is the simple, descriptive quantity, while entropy, according to common misunderstanding, is an abstract, non-descriptive construction.

Disposal:

Draw the *T*(*S*) diagram, not the *S*(*T*) diagram.

Draw the *v*(*p*) diagram, not the *p*(*v*) diagram.

Also:

Draw the *E*(*p*) diagram, not the *E*(*v*) diagram.

Friedrich Herrmann

2

Energy

2.1 Forms of energy

Subject:

It is common knowledge that energy exists in various forms. Kinetic, potential, electric, chemical energy and heat are examples known to everybody; “converting energy from one form into another” is a common way of speaking.

Deficiencies:

Although we often speak about energy forms, we run into difficulties as soon as we try to define them. We are not consistent in the necessary distinction between the forms of stored and transmitted energy. On the contrary, in our casual formulations we tend not to differentiate the two concepts. While for heat and different types of work certain rules have been established, the classification of storage forms of energy seems vague and arbitrary, with the exception of some mechanical textbook examples. Which part of the energy of a steel spring or of an air molecule is mechanical, thermal, chemical, electric or magnetic? Which part is translational, rotational, oscillatory or electronic? Which part is kinetic or potential? Which part is ordered or unordered? The fact that we obtain reasonable results without knowing the answers to these questions leads to the conclusion that the classification is of no importance for our physical arguments.

Origin:

In order to account for the role of energy within the network of physical phenomena, enumerating energy forms is a means of expression which is difficult to avoid. This can be seen in a citation of *F. Mohr* (1837) from the time before the discovery of the conservation of energy: “In addition to the 54 known chemical elements there exists in nature yet another agent, the name of which is Force: Under appropriate circumstances, it appears as movement, chemical affinity, cohesion, electricity, light, heat and magnetism, and from each of these forms of appearance, all of the others can be brought into being.”

Disposal:

We save many words if we refrain from useless differentiations. It is often comfortable to speak about bottle milk and carton milk. It is completely useless, however, to call the process of transferring or drinking it “milk conversion,” or to define the content of a glass or of the stomach as different “forms of milk.” The situation is the same when speaking about the energy. The clearest, but perhaps not the most comfortable solution is to refrain from speaking about energy forms completely. Of course, just as for a patient who after a long period of convalescence leaves his crutches for the first time, it takes time until one is acquainted to the newly acquired freedom and also to be able to cover difficult terrain.

Georg Job

2.2 Pure energy

Subject:

In textbooks and scientific reviews one often finds statements that say electromagnetic radiation is pure energy. Here is an example of such a formulation [1]: “When a positron encounters an electron, the two particles annihilate each other and produce pure energy in the form of gamma radiation.” Or another example [2]: “A massive particle and its anti-particle can annihilate to form energy, and such a pair can be created out of energy.” A similar point of view is expressed in the following formulation [3]: “... light can also be described in terms of photons, discretely emitted quanta of energy.”

Deficiencies:

It is obvious that an electromagnetic wave is not pure energy. The electromagnetic field is a physical system, i.e. a thing, for which every standard physical quantity has a certain value, and not only the energy.

So, in general for an electromagnetic field, apart from just the energy, the extensive quantities momentum, angular momentum and entropy also have non-zero values. But intensive quantities also have certain values, just as is the case for other systems. So the electromagnetic field has a pressure at every point. (The pressure depends on the direction and is therefore a tensor.) In certain states, i.e. in those states that are usually called thermal radiation, the field has a certain temperature and a certain chemical potential.

Identifying the radiation with one single quantity is simply not correct. The radiation is a physical system, something that is given to us by nature. Physical quantities on the contrary are products of the human mind. They are tools for the description of systems.

Correspondingly, a photon, the elementary portion of the system “electromagnetic field”, is more than just a quantum of energy. The photon also carries other extensive quantities in addition to energy, such as momentum and angular momentum.

The confusion between the concepts “quantity” and “system” also manifests in a kind of formulation often encountered in which energy and matter are presented as two concepts on an equal footing [4]: “So if galaxies are all moving away from one another […] it seems logical that they were once crowded together in some dense sea of matter and energy.”

Origin:

There are probably two causes for the erroneous identification of the quantity “energy” and the system “electromagnetic field.” Apparently, on the one hand the energy was seen as more than just a variable in a theory, and on the other hand, the field was not taken seriously as a system.

After the introduction of the energy in the middle of the 19th century, its comprehensive significance in science was quickly understood. However, the enthusiasm about the importance of the new quantity led to an overestimation and misinterpretation of it. Energy was conceived, in particular in the circle of the “energeticists”, as a kind of substance. So, one can read in Ostwald‘s 1908 book *The Energy* [5]: “Therefore, the energy is contained in all real and concrete things as an essential component, which is never absent, and therefore we can say that the energy embodies the actual reality.”

On the other hand, the electromagnetic radiation was not accepted as what we today understand by the concept. We now know that it is a system like other system, for instance an ideal gas, or the phonon system of a solid. Like other systems, the electromagnetic field consists of elementary portions. What the hydrogen molecules are to the hydrogen gas and the phonons are to the lattice system of a solid, the photons are to the electromagnetic field.

This misunderstanding of the physical quantity “energy”, as well as of the physical system “electromagnetic field”, has left its traces. Although we have known better for a long time, we still easily use sentences like those cited at the beginning.

Disposal:

Instead of saying that pure energy is created in a reaction of an electron and a positron, say that a photon results. And instead of saying electromagnetic radiation is pure energy, say that the radiation carries energy, but besides energy it also carries other extensive quantities such as momentum, angular momentum and entropy.

[1] *Scientific American*, December 1993, S. 44

[2] *R. Penrose*: The emperor's new mind, Oxford University Press, S. 308

[3] *Scientific American*, April 1993, S. 26

[4] *Scientific American*, October 1994 S. 32

[5] *W. Ostwald*: Die Energie. – Verlag Johann Ambrosius Barth, Leipzig, 1908, S. 5.

Friedrich Herrmann

2.3 Power

Subject:

The name “power” for the physical quantity P appearing in the equation $P = dW/dt$

Deficiencies:

The equation $P = dW/dt$ refers to a given area or surface. dW is the energy transmitted through this surface. Usually it is called “the work done” by the system at one side of the surface on that at the other side. As a consequence, P is the energy transported through the surface per time interval, or in other words, the energy flow rate, or energy flow for short. If the energy is flowing along a well defined path, and if the energy flow is the same at any cross section of this path, then P can also be attributed to the whole path or conductor.

Thus, P has a simple meaning. However, the denomination “power” does not clearly express this meaning. The word suggests attributing the word “power” to an entire device – an electric motor for instance – instead of to the cable leading to the device. In order to point out that a transport process is meant, sometimes one speaks about the “transmitted power”. This way of speaking is particularly awkward, since what is transmitted is energy, not energy per time.

Origin:

The word “power” for the above mentioned quantity came into being at a time when physics was not yet able to localize either the energy itself, or energy flows. One was aware that the decrease of the energy in one system was related to its increase in another system. However, for one of the most important energy transport processes a distribution of the flow could not be defined, i.e. for the transport of electric energy. That is why the quantity P was used to describe the change of the amount of energy in a system, i.e. in some fixed location. Thus, P was attributed to a body or a device, and not to a cross section.

Disposal:

Don't call the quantity P “power” but “energy flow rate” or “energy current.”

Friedrich Herrmann

2.4 The energy conservation law

Subject:

The formulation of the energy conservation law does not seem to be trivial. The quotations (1) and (2) are taken from school books, and quotation (3) is from a university book.

(1) “The total energy of a body can be distributed among different forms of energy. – Without the transfer of energy to or from other bodies the total energy of the body remains constant”... “If several bodies are involved in the exchange and transformation without friction being present, the sum of kinetic, elastic and gravitational energy remains constant.” … “If friction is taken into account, the internal energy of the bodies and of the environment are part of the energy sum.”

(2) “Theorem of the conservation of mechanical energy: In an energetically isolated system the sum of the mechanical energies remains constant, as long as the mechanical phenomena take place without friction. Energy is never lost, nor does new energy come into existence; it transforms from one mechanical form into another…. According to this theorem there exists a state variable for an energetically isolated system, called mechanical energy, which can appear in different forms, whose value is always conserved. Therefore, the energy of such a system is a conserved physical quantity.”

(3) “Now the energy law can be formulated as follows: The amount of heat ΔQ supplied to a system from the outside serves to increase its internal energy ΔU, e.g. its temperature... or its electrical or chemical energy, and serves to realize the work ΔW, which we will consider negative when it is delivered by the system, so that

$\Delta U = \Delta Q + \Delta W$.”

Deficiencies:

A simple fact is described in such a way that it is hardly possible to recognize its simplicity. One might argue that before formulating the energy theorem, much has to be taken in consideration. However, one should eventually pronounce it in all clarity: Energy cannot be produced or destroyed. And there should be no qualms with this sentence. Otherwise the idea unavoidably comes up that conservation itself is a difficult concept.

Origin:

See the article “isolated systems”.

Disposal:

Formulate energy conservation in the same way as the conservation of electric charge, i.e. without any ifs or buts, for instance as follows: Energy can neither be created nor destroyed.

Friedrich Herrmann

2.5 Where is the energy?

Subject:

The verbal description of energy transport processes and energy balances by statements as the following:

“The mechanical power tells us, how fast work is realized.”

“In a closed system the sum of all energies is constant.”

“The mechanical work that is realized on a body is equal to the change of its energy.”

Deficiencies:

Energy is an extensive quantity for which a conservation law is valid: When the value of the energy in system A decreases, it increases in another system B by the same amount. Since the end of the 19th century it is known that the law is valid not only in this global form. It became possible to define an energy density ρ_E and also the path of the energy when going from A to B, in the form of the energy flow density $\mathbf{j}_E$. Thus, it became possible to say where the energy is and which way it goes from one place to another. Energy conservation could now be expressed by means of a continuity equation:

$$\frac{\partial \rho_E}{\partial t} + \mathrm{div}\ \mathbf{j}_E = 0$$

The statements cited under “subject” are formulated in such a way that they require only the older form of the law of the conservation of energy. In other words: They allow for the idea of an action at a distance: The energy of B can increase and that of A decrease without the intervention or participation of a third system that connects A and B.

The idea of the energy thereby fostered is inconvenient for two reasons: Firstly, It does not reflect the modern point of view that actions at a distance do not exist, and secondly, the verbal description of an energy balance is unnecessarily complicated.

Origin:

Immediately after the introduction of the physical quantity energy into physics around the middle of the 19th century there was no other choice. Energy conservation could be observed only by comparing the energy contents of two systems A and B. It was not yet clear whether a local distribution of the quantity could be defined in every case, and one was still unable to specify a path for the energy going from A to B. Only in the case of heat it was clear what happens physically between A and B but not so for electric or mechanical energy transports.

The local distribution of energy within electric and magnetic fields was then given by Maxwell, but the field energy became a true local quantity only by the work of Heaviside and Poynting in 1884 [1].

However, the issue was not yet resolved generally. In the year 1892 Heinrich Hertz in his *Investigations about the propagation of the electric force* [2] was skeptical: “It seems to me that a major worry resides in the following: Does the idea of the localization of the energy and its flow from point to point make any sense, given our actual knowledge about the quantity? Such considerations have not yet been carried out for the simplest energy exchanges of common mechanical processes; thus, the question remains unanswered if and to what extent the concept of energy allows for such a treatment.” In the same book somewhat later he expresses it with the following words: “If a steam engine drives a dynamo by means of a drive belt which is running back and forth, and the dynamo feeds an arc lamp by means of a cable that runs back and forth, it is common practice and unobjectionable to say that the energy is transmitted by means of the belt from the steam engine to the dynamo and from the dynamo by means of the wire to the lamp. But does it have a clear physical meaning to pretend that the energy moves from point to point through the tended belt against the direction of movement of the belt? And if not, can it make more sense to say that the energy moves within the wires, or –according to Poynting– in the space between the two wires from point to point? The conceptional obscureness that shows up here greatly requires an elucidation.”

However, when Hertz’s book appeared, the obscurity was already elucidated. In 1891 Heaviside [3] described also mechanical energy transmissions locally. The situation became even clearer, in particular for the German reader, when in 1898 Gustav Mie published his comprehensive work “Outline of a general theory of energy transmission” [4]. Here his opinion in his own words: “If between two material systems A and B that are separated in space there are only energy transitions which are related with state variables of the points of a body C that connects A and B, in such a way that it is possible to calculate the energy transition *dE/dt* only by using the state of of all the points of C, then one says that the energy *dE* has been transmitted from A to B through C. […]. Energy transitions, i.e. any changes of the energy distribution in space, can be realized only by energy transmission.”

The sentences cited under “Subject”, which are typical for our actual way of speaking about energy, show, that the language which developed immediately after the introduction of the energy was conserved so as if the work of Poynting, Heaviside and Mie had never been published – to the detriment of all those young people who try to get a clear idea about the quantity energy.

Disposal:

Introduce the energy in such a way that it becomes clear from the beginning that energy is distributed in space, and that it can flow. Formulate energy conservation as follows: “Energy can neither be created nor destroyed.“ The concepts of *work*, *power* and *energy form* are superfluous [5,6].

[1] *J. H. Poynting*: On the transfer of energy in the electromagnetic field, Phil. Trans. A 1884, S. 343-361

[2] *H. Hertz*: Untersuchungen über die Ausbreitung der elektrischen Kraft, Johann Ambrosius Barth, Leipzig 1892, S. 234 und 293

[3] *O. Heaviside*: Electrician 27, 3. Juli 1891

[4] *G. Mie*: Entwurf einer allgemeinen Theorie der Energieübertragung, Sitzungsbericht der mathematisch-naturwissenschaftlichen Classe der kaiserlichen Akademie der Wissenschaften, CVII. Band, Abtheilung II.a, 1898, S. 1113-1181

[5] *G. Job*: Energieformen, Forms of energy, article 2.1

[6] *Herrmann, F.*: Power, article 2.3

Friedrich Herrmann

2.6 Potential energy

Subject:

From an encyclopedia:

"The potential energy is one of the energy forms of physics. It is that energy which a body possesses due to its position in a force field (for instance a gravitational or an electric field)."

From a school book:

"Example: The potential energy of a satellite…"

Deficiencies:

In the citations the potential energy is attributed to a body.

If one is convinced that energy can be localized (and this is the general conviction of physics since the end of the 19th century), then one will understand the citations as follows: Bodies contain potential energy. And as a consequence: The potential energy must be distributed within the bodies in a well-defined manner. These conclusions, however, would not be correct. The potential energy is not contained in the bodies but in the fields that are mainly situated between the bodies.

In particular our second citation shows that something cannot be correct. If the potential energy is attributed to and thus localized within the satellite, then the potential energy of the system earth-moon would be localized within the moon, and when finally considering a binary star system composed of two stars of equal mass the potential energy would be localized in only one of the stars – which cannot be true for symmetry reasons. Sometimes the term potential energy is also used when the momentum transfer between two bodies goes not via a field but by means of an elastic spring. In this case usually the energy is correctly attributed to the spring. However, the name "potential" for the energy is not convenient. According to the Merriam-Wester dictionary, the adjective "potential" means: "*existing in possibility; capable of development into actuality*". This definition does not agree with the energy that is stored in a spring. Just as the kinetic energy is contained within a moving body, the energy that is supplied to a spring when expanding it is stored within the spring. For both of them a density distribution can be indicated (i.e. the energy can be localized) and both can (in principle) be measured by the relativistic mass increase.

Origin:

The inconvenient wording seems to have several causes or origins.

1. The concept potential energy stems from a time when energy could not yet be localized (before 1890).

2. When teaching, the concept is usually introduced by considering a small body in the gravitational field of the earth. In this case the potential energy can be calculated by means of the equation $E = m \cdot g \cdot h$. Here h is the height of the small body with respect to a zero point that is firmly connected with the earth. h does not appear as the distance between two bodies, namely the small body and the earth, and h does not appear as the height of the earth above the small body, but that of the small body above the earth or the surface of the earth.

3. Often, the context in which the term is employed is the movement of two bodies that interact gravitationally, and where the mass of one of the bodies is much larger than that of the other. Let us consider the famous falling apple and begin with the momentum balance (in the center of mass system): Only the earth and the apple participate in the process; the momentum that the apple is gaining is lost by the earth. The momentum of the field is almost zero and does almost not contribute to the momentum balance. Things are different when we consider energy. The kinetic energy of the earth does practically not change (since the mass of the earth is much greater than the mass of the apple). The energy which the apple is receiving does not come from the earth but from the gravitational field.

The same is true when two bodies are coupled by a spring instead of a field. Here too, the momentum exchange is between the two bodies, whereas the energy exchange is between the light body and the spring.

Disposal:

Small solution: Avoid formulations that attribute the potential energy to a body. Here is an example from another school book:

"The potential energy of the system 'earth - body of mass m' with respect to a reference level, that can be arbitrarily chosen, is…"

This wording is better than that of our initial citations. However, it still suggests the idea of actions at a distance, since the system is called "earth - body of mass m". The field as a part of the total system is not mentioned.

Great solution: Introduce the field from the very beginning as the third partner and say clearly where the energy is localized, namely within the field.

The adjective potential should be avoided in any case.

Friedrich Herrmann

2.7 Perpetual motion and the energy conservation law

Subject:

Brockhaus* 1839 [1]: “Perpetual motion machine: a machine which, thanks to the driving force which is generated by it, would remain in a steady motion, but whose realization is now thought to be impossible, since well-known laws of nature speak against it. In former times, together with the philosopher’s stone, the elixir of life etc. It belonged to the things on which charlatans were preening themselves and whose discovery was the ambition of mechanicians and mathematicians. ”

Brockhaus 1910 [2]: “Perpetuum mobile (latin), a body that moves incessantly, in particular an aspired device that is shown to be impossible due to the law of the conservation of energy, which would be able to renew its force thanks to its own movement.“

Brockhaus 1953 [3]: “Perpetuum mobile of the first kind, a machine which supplies energy steadily without the need of any work; is in contradiction to the empirical law of the conservation of energy.”

A publication of the Federal patent office from 1985 [4]: “the federal patent court points to the ‘energy conservation law which is recognized and unrefuted in the whole of natural sciences’ according to which ‘energy cannot be created or destroyed in any physical process’, but it can only ‘be converted from one form into another’.

Deficiencies:

The assertion that a perpetual motion machine (PM) of the first kind cannot work because it would violate the energy conservation law, falls somewhat short of the mark.

Imagine you don’t know the energy conservation law and you would like to prove that a perpetual motion machine that has been proposed by someone, cannot work, without trying it experimentally. It will be easy to provide evidence, since apart from the energy conservation law there are always other laws which are also violated: other conservation laws, Maxwell’s equations, the Law of Gravitation etc. Mechanical perpetual motion machines usually violate Newton’s laws, i.e. the law of the conservation of momentum, or they violate the law of the conservation of angular momentum.

The well-known discussion among physicists about why a certain smart proposal of a perpetuum mobile doesn’t work also show that energy conservation is not the only obstacle. Although the discussants know perfectly that energy conservation is violated they consider the refutation satisfying only when yet another reason is found, i.e. the violation of another physical law.

Indeed the energy conservation law is a practical tool for showing that a certain process cannot occur. But it does not play a distinguished role in our context.

Origin:

Since it is the stated objective of the PM inventors to violate the energy conservation law, it is suitable to argue with this law in order to refute the realizability of such a machine. Apparently, PM inventors, who can be found even at the present time, have not too much fantasy. They only focus on devices that violate energy conservation. The reason might be that they consider energy a precious merchandise. They seem not to understand, that they can make just as much money by violating any other law of physics.

Disposal:

Perpetual motion machines (which do not work) are nice and profitable subject for physical discussions. One should not dismiss the theme by saying that the law of energy conservation is violated. Then the impression will result that the structure of physics is such that one can imagine a world in which the laws of physics are the same as in our world except for any one law which has been replaced with another one.

**Brockhaus* is a time-honored German encyclopedia.

[1] Bilder-Conversations-Lexikon, F. A. Brockhaus, Leipzig, 1839

[2] Brockhaus’ Kleines Konversations-Lexikon, F. A. Brockhaus, Leipzig, 1910

[3] Brockhaus ABC der Naturwissenschaft und Technik, VEB F. A. Brockhaus Verlag, Leipzig, 1953

Friedrich Herrmann

2.8 Isolated systems

Subject:

In order to formulate the conservation of energy or of other physical quantities, we often refer to an isolated system. We imagine a region of space whose boundaries are impermeable for a current of the quantity under consideration. The quotations (1) and (2), which refer to the conservation of energy, are taken from books for the secondary high school and are highlighted in these books.

(1) “In a thermally and mechanically isolated system the total energy is constant.”
(2) “In an isolated system the sum of all energies is always constant. The total energy is conserved.

$$E_{\text{total}} = E_1 + E_2 + \ldots + E_n = \sum_{i=1}^{n} E_i = \text{constant}$$

$E_1, E_2, \ldots E_n$ different energy forms”

Deficiencies:

The concept of conservation of an extensive or substance-like quantity is not a difficult concept. This has to do with the fact that we can easily represent these quantities pictorially: We imagine them as a kind of fluid or stuff. The conservation of a quantity X can then be stated in the following way: “X cannot be produced and cannot be destroyed.”

Here the exact wording doesn‘t matter. Conservation is something that we can easily express with words of the common language.

A consequence of this statement is that the value of X in a region of space can change only if a current of X flows into or out of the region. Mathematically the statement can be expressed in the following way:

$$\frac{dX}{dt} + I_X = 0.$$

Here dX/dt is the rate of change of X in the considered region and I_X is the flow of X through the boundary surface.

A formulation of the principle of energy conservation that refers to an isolated system is a special case of this statement. “The system is isolated” means that there is no flow through the boundary surface. However, the isolation is an unnecessary restriction because the considered quantities are conserved independent of whether the system is closed or not.

To convince myself that the number of my students “is conserved”, there is no need to close the door of the classroom. There is no problem if, from time to time, somebody comes in or goes out, as long as I ascertain that the number of students in the classroom increases by one when someone comes in, and decreases by one when someone goes out.

Origin:

The fact that we formulate conservation with reference to an isolated system is a leftover of the troublesome development of the concept of energy as a substance-like quantity. Until shortly before the beginning of the 20th century, the localizability of energy was not acknowledged. It was not yet possible to associate a density, a current and a current density with it. In 1887 Max Planck [1] wrote in a historical survey about the energy:

“... according to this definition the amount of the energy is measured only by these external effects, and if one wants to attribute any imaginary material substrate to the energy, then one has to look for it in the environment of the system; only here the energy finds its explanation and therefore also its conceptual existence. As long as one abstracts completely from the external effect of a material system, one cannot speak about its energy, since it then is not defined... On the other hand, we see from the form of the principle as derived formerly that the energy of a system remains constant, if a process carried out with the system does not cause any external effect whatever the internal effects may be. This observation leads us to conceive the energy contained in a system as a quantity existing independently of the external effects.” And later: “Meanwhile it is unmistakable... that with this substance-like interpretation of the energy we get not only an increase in the conceptual clearness but also a direct progress in the comprehension... However, as soon as one enters into this question, the uncertainty, which lay before in the concept itself, takes upon the form of a physical problem which in principle can be solved...”

This solution came a few years later by Gustav Mie [2]. He showed that the principle of energy conservation can be formulated locally, namely in the form of a continuity equation. From then on, the strange separation of the system and the effects that can be observed only in the environment was no longer necessary.

Thus, it took about 50 years to prove the substance-like nature of energy. However, the expectation that the quantity had this property was there from the beginning: Ostwald [3] in his 1908 booklet, *The Energy*, praised the work of Robert Mayer with the following words: “For our general investigation the essential result of Mayer‘s work is the substance-like view of what he calls force, i.e. the energy. For him this was a well-defined entity; the indestructibility and unproducibility are characteristic for its reality.”

Disposal:

We state the conservation law of the substance-like quantity X in the following way: “Energy, momentum, angular momentum, electric charge … cannot be produced and cannot be destroyed.”

Just as important are statements about the non-conservation of a substance-like quantity, for example: “Entropy can be produced but cannot be destroyed.”

[1] *M. Planck*: Das Prinzip der Erhaltung der Energie. B. G. Teubner, Leipzig, 1908, S. 115.

[2] *G. Mie*: Entwurf einer allgemeinen Theorie der Energieübertragung. Sitzungsberichte der Kaiserlichen Akademie der Wissenschaften. CVII. Band, VIII. Heft, 1898, S. 1113.

[3] *W. Ostwald*: Die Energie. Verlag Johann Ambrosius Barth, Leipzig, 1908, S. 59.

Friedrich Herrmann

2.9 Relaesed energy

Subject:

Energy is released, in French “libérée”, in German “freigesetzt”.

Deficiencies:

It is often said that energy is released. I confess that I don’t really know what that means. Let us consider a statement like this:

- The energy is released by the emission of a photon.

Well, the photon flies away, free like a bird that is released from its cage, or a prisoner from prison. It seems to make sense that the energy is released. But shouldn’t we be consistent and say that the energy is captured or trapped again during an absorption process? But nobody says that. So the fact that the energy flies away cannot be the criterion for being “released”.

Another assumption might be: One is interested in a process in which energy is supplied by a system; it comes out of the system. One is only interested in the system that provides the energy, but not in what happens to the energy afterwards. It doesn’t matter whether something is excited, heated, chemically transformed, vaporized, compressed or accelerated. The following quotation speaks for this interpretation:

- Chemical reactions, in which energy is released in the form of heat, are called exothermic reactions. The released energy can be used, for example, to emit heat and light or to do electrical or mechanical work.

However, statements like the following speak against this interpretation:

- The energy released during braking is recovered in the form of electrical energy, stored in batteries and [...] reused for propulsion via an electric motor.

Here it is clearly stated what happens to the energy emitted during braking. So, is it perhaps “trapped” in the battery again? After all, it will certainly be “released” again when it leaves the battery and is “reused for propulsion” by the electric motor. Or is it that it is always released and released again, thus becoming freer and freer?

How the phrase “energy is released” is used is nicely demonstrated when one enters the word combination for instance at the website Linguee. Linguee spits out dozens of quotations. You can also try the French or German counterpart. What you will notice in any case is that if you replace the “release” by “deliver”, all sentences you find remain clear and correct.

Another remark: When dealing with substance-like (extensive) quantities, one has a great deal of freedom in the choice of words. We can say, for example, that electric charge is *stored*, *distributed*, *concentrated*, it can *flow*, *come* and *go* and *disperse*. All these ways of speaking can be used with benefit in connection with all extensive quantities. But let’s try *releasing*. Would we say we release electric charge when a charged body discharges, or we release momentum when a car brakes, or the bathtub releases water, when it is emptied?

Origin:

Probably a remnant from the time when the quantities energy flow and energy flow density did not yet exist, and when no local balance could be established for the energy, thus roughly from the time before 1900.

Disposal:

A simple rule that I pass on to each of my student teachers: Talk about energy as if it were a substance. Always asks the questions: “Where is it?”, “Where does it come from?”, “Where is it going?”

And if you ever believe you’re in a situation that justifies using the metaphor of being released, don’t forget to say that the energy was locked up before, and that it will be trapped again afterwards.

Friedrich Herrmann

2.10 Useful and wasted energy

Subject:

In publications of the energy economy, of governmental institutions and of universities one can find so-called energy flow charts [1, 2, 3, 4]. They represent the energy balance of a national economy. Such charts show with which energy carriers primary energy enters the economic system, which parts are transformed in other “energy forms”, how great are the losses in this process and in which forms the energy leaves the system. At the exit one distinguishes between useful energy and wasted energy.

Deficiencies:

The following impression is caused by such diagrams: The costumer needs energy in a certain form. That is why the energy has to be transformed and in the process some of it is lost or wasted. One tries to keep the losses as low as possible but a considerable loss is unavoidable in principle, for physical reasons. Once the energy has arrived at the costumer it can be used for what it is really needed.

This view on things does not really hit the mark. We can see it when we realize that every energy loss is due to entropy production. Produced entropy has to be carried away into the environment, and for that purpose energy is needed:

$P_\mathrm{L} = T_0 \cdot I_S$.

Here, P_L is the flow of the lost energy, T_0 is the ambient temperature and I_S is the flow of the entropy that has to be carried away. Two things follow from this observation:

1. From the point of view of physics transformation losses are not unavoidable. Any process can be carried out reversibly. It may be impossible or inconvenient for technical or economical reasons, but physics does not forbid it. Take as an example the “transformation” of the chemical energy of coal (+ oxygen) into electric energy. Here, usually the Carnot factor is hold responsible for the low efficiency. But the process can in principle be carried out in a combustion cell that works reversibly. Thus, the energy that enters the chart and which is called primary energy can be considered useful energy.

2. All of the energy which reaches the costumer ends up in the elimination of produced entropy and is thus wasted. A 100 % of the “useful” energy is finally wasted. And by the way: Also for the costumer the rule holds: Every process can be replaced with a reversible process.

We do not want to say that the above-mentioned energy flow charts are incorrect; neither do we want to say that they are useless. We only believe that they give us the wrong message. It is not true that a well-defined fraction of the primary energy is really “used”. Instead 100 % of the primary energy ends after a many intermediate steps in the entropy production and elimination. The energy flow charts show only the first part of these processes.

Origin:

Why do the energy flow charts end at a certain point? Why don’t they show that all the energy ends up as thermal waste? Because they are created by institutions with certain interests. For the energy companies the charts ends at the fare stage; where the suppliers hold up their hands. The losses before this borderline are the subject of their efforts. What the costumer does with the energy does not matter for them.

Disposal:

Explain the students that all of the primary energy eventually end up in the thermal deposit. There is no physical limit for avoiding entropy production and thus energy loss. In principle every process can be carried out without entropy production. When discussing the technical problems that arise when trying to reduce entropy production students can learn a lot of good physics (and chemistry).

[1] http://www.zw-jena.de/kkimages/energieflussbild1995.gif
[2] http://www.bpb.de/files/WQ93Q3
[3] http://www.ag-energiebilanzen.de/viewpage.php?idpage=64
[4] http://www.energyliterracy.com/?p=293

Friedrich Herrmann

3

Electricity and Magnetism

3.1 Excess and deficit of electrons

Subject:

The terminals of power supplies and batteries are marked with a plus and a minus sign. When discussing simple electric circuits it is often said, that at the minus terminal there is an excess of electrons and at the plus terminal there is an electron deficit.

Deficiencies:

Here we have to do with two incongruities, which are related to one another. We shall show that

– it is awkward to tag the terminals with a plus and a minus sign;

– it is often awkward, and sometimes incorrect to say that at the terminal at lower potential there is an excess of electrons and at the high potential terminal there is a deficit.

The plus and the minus sign suggest, that some physical quantity has a positive or negative value at the corresponding terminal. Does such a quantity exist?

One might think at the electric charge. Let us first ask for the amount of charge sitting on the terminals of a battery including the respective electrode. Its value depends on the capacitance C_B of the battery. We thus treat the battery as a capacitor. The electric charge on the terminals, together with the electrodes and the internal conductors is

$$Q = C_B \cdot U_B \,, \qquad (1)$$

where U_B is the voltage of the battery. Q would be the charge of the terminal (and electrode) if the midpoint between the electrodes would be at Earth potential. The electric potential of the plus electrode would then be by $U_B/2$ above and the minus electrode by $U_B/2$ below Earth potential. However, this is a special case that is almost never realized. In general the average potential of the battery will be different from the Earth potential and thus the battery will carry a net charge, whose counter charge sits at the Earth. The net charge is then:

$$Q = C_{plus} \cdot U_{plus} + C_{minus} \cdot U_{minus} \,,$$

where C_{plus} and C_{minus} are the capacitances of the plus and minus terminals against the Earth, and U_{plus} and U_{minus} are the voltages between the terminals and the Earth. The capacitances C_{plus} and C_{minus} are typically of the same order as C_B, whereas U_{plus} and U_{minus} depend on the circuit as a whole; it may be grounded somewhere but it may also be carried at a high positive or negative electric potential. Thus, in general one cannot say that the plus terminal carries positive and the minus terminal negative electric charge.

These considerations show that the plus and minus sign at the batterie's terminals neither correspond to the electric potential. The plus terminal must not be at a positive potential and the minus terminal must not be at negative potential. And there is no other quantity that would be correctly characterized by the plus and the minus sign.

There is no doubt that this designation is the cause of incorrect conclusions.

It also follows that the claim that there is an excess of electrons at one terminal and a deficit at the other cannot be generally correct. A student will believe that such a statement means that the plus terminal is not electrically neutral but that it carries positive charge. We just have seen that this must not be true.

But even if we arrange the electric potentials in such a way that the plus terminal's potential is positive and the minus terminal's negative (when Earth potential is defined as zero), so that the plus terminal is positively charged and the minus terminal negatively, even now it would be out of place to characterize the terminals by speaking of an electron excess or deficit.

The capacity C_B in equation (1) is of the order of 10^{-10} F. Since a typical voltage is 1 V the excess charge is of the order of 10^{-10} C. When mentioning an excess or deficit of electrons one suggests that this has something to do with the flow of charge or electrons that flow in the circuit under typical conditions. However, the charge that is crossing a section of the wire of the circuit with a light bulb in one second is greater by 10 orders of magnitude.
The inappropriateness of the argument can best be seen when comparing the battery in the electric circuit with a water pump in a water circuit. A battery without a load would correspond to a water pump that is filled with water but with its inlet and outlet blocked. Nobody would characterize the inlet and outlet by saying there is an excess and a deficit of water. The slight excess and deficit that actually exists is due to the non-zero compressibility of the water. But we see immediately that this excess is not a necessary condition for the operating of the pump. The pump would do just as well with a liquid of zero compressibility.

Origin:

Most of the subjects of our column concern concepts or descriptions that in a former time have been justified. Here we have an example of an incongruity that was an incongruity from the beginning.

Disposal:

Characterize the terminals of a battery or a power supply with “high” (H) and “low” (L) instead of plus and minus. This is common use among electronic engineers. This labeling refers to the electric potential. Another possibility would be to label them “out” an “in”, which refers to the electric charge, and not to the electrons.

Friedrich Herrmann

3.2 Two types of electric charge

Subject:

Electric charge comes in two types, called positive and negative. Like charges repel and opposites attract.

Deficiencies:

Since electric charge is a physical quantity, the above wording suggests that there are two of them. Let us call them Q_A and Q_B. We can indeed describe the electric state of a body by indicating how much of Q_A and Q_B it contains. However, the quantities Q_A and Q_B have an unpleasant property: Each of them taken separately is not a conserved quantity. However, the production and annihilation of them are coupled: The production of Q_A is accompanied by an equal production of Q_B. Mathematically and conceptually it is simpler to use one single quantity "electric charge", which can admit positive and negative values. Obviously, for this quantity a conservation principle is valid.

There is even more of a muddle in the case of magnetic poles. Whereas in the electric case the terms "positive" and "negative" suggest the mathematical relationship between the two "types" of charge, the denotations "north" and "south" do not suggest at all, that the strength of a magnetic pole can be described by a single extensive quantity. The names "north" and "south" suggest that the poles of a magnet have different qualities, between which there is no transition, similar to the properties "male" and "female" of persons or animals.

Origin:

When electrostatic phenomena were discovered the question was if there are two distinct electric fluida or only one fluidum. The two-fluida theory has left its remnants until this day.

Disposal:

Avoid speaking of types of electricity. There is only one physical quantity electric charge, which can admit positive and negative values. Call the poles of a magnet the positive and the negative pole. Instead of speaking of like and opposite charges (magnetic poles), call them electric or magnetic charges of the same or opposite sign, respectively.

Friedrich Herrmann

3.3 The conventional flow notation

Subject:

The direction which is attributed to an electric current is based on a convention. Before the true direction of electron flow was discovered the direction of the electric current had been defined in such a way that the current flows from the plus to the minus terminal of a power supply (in the external part of the circuit).

Deficiencies:

When asking for the direction of an electric current, one is asking for the orientation of a vector. The vector that characterizes the direction of the electric current is the current density vector, just as the energy flow density vector characterizes the flow direction of the energy, or the mass flow density vector tells us the direction of a mass flow. Now, the direction of the electric current density vector does not depend on a convention. It follows from the continuity equation for the electric charge, which relates the charge density ρ to the current density $\boldsymbol{j}$:

$$\frac{\partial \rho}{\partial t} + \text{div}\ \boldsymbol{j} = 0$$

The equation tells us, that the charge density ρ decreases at a given place if the divergence of $\boldsymbol{j}$ is positive at the same place. In other words: The electric charge decreases in a small region if an electric current is flowing out of this region. This statement is analogue to the following: The amount of water in a container decreases if a water current is flowing out of the container.

We see that the orientation of the current density vector is defined, as soon as we have disposed of the sign of the electric charge. We could indeed redefine the direction of the electric current, but only by redefining the sign of the electric charge. If we want to keep the minus sign for electrons and the plus sign for protons, then there is no choice for the direction of the electric current.

Origin:

When it is claimed that the direction of the electric current is based on a convention, what is meant is not the direction of the current density vector $\boldsymbol{j}$ but the direction of motion of the mobile charge carriers, i.e the vector of the drift velocity $\boldsymbol{v}$ of the carriers. Both vector quantities are related by

$\boldsymbol{j} = \rho \cdot \boldsymbol{v}$.

It does not matter if positive charge carriers move in one direction or negative carriers in the other – the current density direction is the same.

Since the direction of $\boldsymbol{v}$ is the same as that of the mass current density or the particle current density of the charge carriers, one can diagnose, that the charge current is mistaken for the mass or the particle current.

Disposal:

Distinguish thoroughly between the concepts charge and charge carrier. Distinguish also between two directions: the flow direction of the electric charge and the direction of motion of the charge carriers (or the direction of the mass current density vector). Whereas the electric charge flows (outside of the battery or power supply) from high to low potential, the charge carriers move in one or the other direction depending on the sign of their charge.

To make the distinction clear in the class room I carry out the following experiment: Pupils sitting in a row pass red and blue tokens to each one's neighbor. We imagine that each red token is 10 euros worth, and each blue one minus 10 euros. Every pupil, except the two at the ends, owns one red and one blue token, i.e. his monetary property is zero. Those at the ends owe a great number of tokens. We now realize several money value transports from the leftmost to the rightmost pupil. A metronome is beating, and at each beat each pupil passes a token to his neighbor. The first transport is as follows: At each metronome beat each pupil - except the pupil at the right end of the row - passes a red token to his neighbor at the right. Thereby each pupil remains with his monetary value zero except the two at the ends: The leftmost gets poorer and poorer, the rightmost richer and richer. Next we realize the transport of monetary value from left to right in another way: On each metronome beat each pupil passes a blue token to his left neighbor. Again all the pupils of the chain remain with zero euros except those at the two ends, and again the one the left end gets poorer and the one at the right end richer. A third possibility for a value transport from left to right is that each pupil passes one red token to his neighbor at the right and simultaneously a blue one to his left neighbor. In each of the three transports the monetary value goes from left to right, whereas the "value carriers", i.e. the tokens, move in one or the other direction.

Friedrich Herrmann

3.4 The current and its article

Subject:

The following formulations, in which the word “current” appears without an article, are taken from physics text books and from the internet: “Conventional current assumes that current flows out of the positive terminal of a power supply.” “When a potential difference is applied to a resistive element, current flows according to Ohm’s Law…” “The flow of water through a system of pipes can be used to understand the flow of current through an electric circuit.”

Deficiencies:

By an electric current we understand in physics the flow of electric charge through a conductive medium. Thus the term *electric current* describes a phenomenon.

Possibly the reader may not have noted anything objectionable in the foregoing citations. But let us remember a rule of grammar: No article is used before uncount nouns when talking about them generally. Among the uncount nouns are all substances:

“The ring is made of gold”, “Water flows downhill”, “Hydrogen reacts with oxigen forming water”, “Hot air rises”, “I need money because I want to buy wine”.

In all of our initial citations the noun “current” is used without an article. If by current we really mean a phenomenon, then we have to use it with an article. We must talk about the electric current in the same way as we talk about a water current or a stream of people or a money flow, i.e. always with an article.

If, as in our initial citations I drop the article, then the meaning of the statements clearly is: I am talking about a kind of stuff, like wine, gold or money.

Particularly interesting is the expression “current flow”, which is often found even in the scientific literature. Compare with “water flow” which means flowing water. So current flow would mean flowing current. The current seems to be something that can, but must not flow. Again we see, that the word current is used for a substance-like entity. It is easy to identify what is meant by current (without an article): the electric charge. The citations become correct when replacing the word current by electric charge.

Origin:

Probably simply the unmindful dealing of experts with the language. Their concern is not to preserve and cultivate the language.

Disposal:

1. When teaching physics, use the word current only to denote a phenomenon and thus use it only with an article. One may comply to the habit of the experts by allowing the use of the word *current* or *electric current* as a name for the quantity *I*, i.e. the electric current intensity. But if we do so we should tell to our students explicitly, that we use the same word with two different meanings: as a name of a phenomenon and as a name of a physical quantity.

When we discuss anything that is flowing, tell to the student from the beginning what it is that flows: water, electric charge, energy… Never say that current is flowing.

Friedrich Herrmann

3.5 Test charge

Subject:

The electric field strength is usually defined by the equation $\vec{E} = \frac{\vec{F}}{Q}$:

"By field strength we mean the ratio of the force acting on a charge to the amount of this charge."

Some authors believe to be more careful, when defining:

$$\vec{E} = \lim_{Q \to 0} \frac{\vec{F}}{Q} .$$

Deficiencies:

When introducing the electric field strength by means of the equation

$$F = Q \cdot E \qquad (1)$$

one may have two intentions. First, one wants to present a procedure that allows to determine the values of the field strength and, second, one wants to create and foster an intuitive idea about the physical system "field". We believe that the introduction of E via equation (1) is not convenient for one or the other purpose.

1. I cannot remember that in my long life as a physicist I ever determined an electric field strength by using equation (1) – which does not mean that I never had to do with electric field strengths. On the contrary, I have calculated and measured field strengths many times – but not by employing equation (1). The electric forces that enter in equation (1) are, on the macroscopic scale, so minuscule that one would not be very happy with such a measurement.

But how are electric field strengths determined practically? In principle any equation which contains the quantity E can be used. In the case of the homogeneous field of a plate capacitor it is particularly convenient to use the relation $E = U/d$. Another workable equation is

$$\sigma = \frac{\varepsilon_0}{2} \vec{E}^2$$

The mechanical stress or momentum flow density σ is obtained from force (momentum current) divided by surface area. When dealing with a capacitor one has to measure the force that one plate exerts on the other.

Another problem with the measurement of E via $F = Q \cdot E$ is that the procedure is not really transparent. Indeed, the field strength that one measures has not the same value as that of the field which is present at the position of the test charge. This is the reason why some authors prefer to say that the "test charge" has to be small, or even tend to zero. It is believed that the measuring process thereby becomes conceptually clearer. Indeed, it seems plausible that the test charge should be small. We all know from other kinds of measurements that the measuring instrument should not disturb the system on which the measurement is carried out. (A voltmeter must have a high internal resistance, a thermometer a low heat capacity.) In our case however, a high value of the charge of the test body does not falsify the result of the measurement of E (as long as no electrostatic induction occurs, i.e. as long as all charged particles are fixed in space). Actually, the test charge can have any value. This value can even be much greater than that of the charges which generate the field, without falsifying the measuring result.

Thus, the test charge must not be small. If it is chosen to be small the problem arises that the force, which is small anyway is getting even smaller.

2. If the mental representation of the electric field is based on equation (1) then the idea may result, that the force on a test charge is the only property of the field. This however would be a bad starting point for learning electrodynamics, since the final goal is to imagine fields as physical systems in its own right. We must not forget that most of electromagnetic goings-on in the world is not related to electric charge. Charge only tells us about the interaction between electromagnetic fields and matter.

Origin:

The great master Maxwell did so – on one of the first pages of his 1000 pages work. It may have appeared natural to Maxwell, since his intention was to explain the whole of electrodynamics mechanically.

Disposal:

Do not couple the idea of the electric field in the first place to the force on a test charge. Introduce the field as a discrete system with various properties. Show that it is an energy store. Next show that it can be characterized by means of a single vectorial quantity. Only then introduce procedures to measure the field strength, but not only by means of the force on a test charge.

Friedrich Herrmann

3.6 Where is the field?

Subject:

To represent a field graphically one almost exclusively uses field line pictures.

Deficiencies:

Field lines are somewhat misleading when we ask the question of where a field is located. Since the field is distributed in space and the field strength changes from point to point, one might argue that this question does not make sense. But it does. When we ask about where is the air of the atmosphere of the earth, we know very well how to answer: by specifying the density distribution of the air. Qualitatively we might say: there is much air at the bottom, few at high altitude and almost none above 40 km.

The only quantity of a field that has the character of a density is the energy density. So, if we want to get an idea about where the field is, or how it is distributed in space, it is reasonable to ask for the energy density. However, the field line picture does not give an adequate information about the energy density, since we read it intuitively as a stream line diagram.

Consider an electrically charged sphere with radius R. The flux of the electric field strength far away, i.e. for a great distance r from the center, is the same as for a small value of r. Thus, the field line picture suggests that in a volume element of thickness dr far away out there is the same amount of field as in in element of the same thickness dr further inside.

Such a conclusion would be correct for a radiating object like a star, if the energy density is taken as a measure of how much field there is: the energy is the same in every volume element of thickness dr.

For a static electric field however, such a conclusion would not be correct. The electric field strength decreases with the second power of r, the energy density with the forth power. Therefore, 90 % of the field energy is located within a sphere of radius 10 R, 99 % within a sphere of 100 R. One can say, that in this sense the field is located in a relative small region around the charged sphere.

We consider yet another example: the magnetic field of a solenoid. The field line picture suggests, that the field within the solenoid is more concentrated but that a considerable part of the “amount of field” is located outside of it. Again, the impression is very different when considering the energy distribution. If the solenoid is not too short, almost all of the energy is located within the solenoid – in the same way as almost the whole energy of the electric field of a capacitor is located between the plates of the capacitor.

Origin:

Usually, the field is defined as a region of space in which forces are acting. These forces can be recognized in the field line picture. As a consequence, the field lines are the only concrete anchor for an intuitive idea or a mental representation of the field.

Disposal:

Introduce the field as an autonomous system, i.e. not only as a mathematical tool for calculating forces. Since a field is an extended system we can represent it by a density distribution, even before showing that forces are exerted on a body which is brought into the field. Only thereafter we show that the “material” of the field is anisotropic. We proceed in the same way as we would do when explaining to somebody what we understand by the material “wood”. We would not begin by drawing lines which express the texture of the material. We rather begin by saying that wood is a homogeneous material with a certain density.

Friedrich Herrmann

3.7 The hysteresis curve

Subject:

When the subject “Magnetic fields in matter” is discussed, dia-, para- and ferromagnetism is introduced. For ferromagnetic substances the hysteresis effect is characteristic. Among other things, the concept of remanent magnetism is introduced.

Deficiencies:

Although magnetic forces are more pronounced and more easily accessible by simple experiments than electrostatic forces, physics students as well as physics teachers, are less versed with magnetostatic phenomena than with electrostatics. One of the reasons for this deficiency is the tradition of explaining ferromagnetism by introducing the hysteresis effect. As a result, the learner gets the impression that the behavior of magnets is essentially determined by the complicated hysteresis curve.

Actually, the hysteresis curve can be considered a manifestation of the imperfection of a magnetic material. We consider two extreme cases of magnetic materials: the perfect soft magnetic and the perfect hard magnetic substances. A perfect soft magnet does not allow for a magnetic field in its interior. Inside of such a material $\mathbf{H} = \mathbf{0}$. Thus, a soft magnetic material is analogous to an electric conductor with respect to an electric field: An electric conductor does not tolerate an electric field in its interior, we have $\mathbf{E} = \mathbf{0}$. A perfect hard magnetic material is characterized by the property, that its magnetization cannot be changed by means of a magnetic field. Thus, $\mathbf{M}$ = const. It is exactly this property which is wanted. A “permanent” magnet that changes its magnetization under the influence of a magnetic field is not a permanent magnet. Actually, both types of perfect materials can nowadays be realized to within a good approximation. The hysteresis curve expresses that by means of an external field which is sufficiently strong, one can destroy the permanent magnetization of a hard magnet, or one can reach saturation of a soft magnetic material. Under normal conditions, however, these phenomena will not be significant. Thus, beginning the introduction of the magnetism of materials with the hysteresis curve means to begin with imperfect materials. It is similar to beginning the discussion of elastic springs by overstretching the spring. Also in this case a hysteresis effect is observed.

Origin:

Only some decades ago it was appropriate to introduce ferromagnetism via the hysteresis curve. The materials which could be realized at that time were still far away from being perfect hard or soft magnetic materials. It was easy to change the magnetization of a permanent magnet. When the geometry of a magnet was unsuitable, the magnetization succumbed the magnet's own field. Under these circumstances it was appropriate to speak about a remaining or remanent magnetization.

Disposal:

We begin the discussion of magnetism in matter with the introduction of the perfect hard magnetic and the perfect soft magnetic material. For a hard magnetic material we have $\mathbf{M}$ = const, and for a soft magnetic material we have $\mathbf{H} = \mathbf{0}$. The magnetization of a permanent magnet is not disdainfully called remanent magnetization. The hysteresis curve, as well as dia- and paramagnetism are introduced in the solid state physics course at the university.

Friedrich Herrmann

3.8 The field as a region of space with properties

Subject:

Physicists consider the concept of field a difficult concept. When reading textbooks, one gets the impression that it is almost a mysterious entity. The following citations are from different sources:

“The attraction…is independent of the intermediate matter and takes place even in empty space! This endues the space around a magnet with a particular significance; it is called a magnetic field.”

“…magnetic field, region in the neighborhood of a magnet…”, “… there is something rather strange about the space surrounding a charged object…”

“Empty space becomes the carrier of a physical property. Such a space is called a field.“

“field, in physics, region throughout which a force may be exerted;…”

Deficiencies:

A field is a physical system that does not differ fundamentally from other systems, such as an ideal gas, a rigid body or a perfect fluid. Like for other systems, the quantities energy, momentum, angular momentum and entropy have well-defined values. Like other “material” systems, it has a pressure and it may, depending on its state, have a temperature. Like other systems, it consists of elementary portions, in the case of the electromagnetic field the photons.

It is therefore justified to view a field a concrete entity, just as a material system, such as air or water, for example.

In the definitions quoted above the field is termed a “region” or a space”. Pupils and students imagine space as empty. Now they learn that the empty space has properties. There is nothing, but this “nothing” has properties. No wonder, that field is perceived as a difficult concept.

Origin:

For Faraday, the inventor of the field concept, the field concept was simple. It did not make great demands on our capacity for abstraction. For him and his contemporaries space was filled with a medium, the “ether”, about which one had a fairly concrete idea. Fields were no less concrete structures: they were areas of the ether in a particular altered state. A characteristic of this state was that the ether was under mechanical stress.

Maxwell, who further developed Faraday's ideas and gave them a mathematical form, defined the field as follows:

“The Electric Field is the portion of space in the neighbourhood of electrified bodies, considered with reference to electric phenomena.” [1]. Notice that for Maxwell the whole space was filled with ether, thus speaking of the space was the same as speaking of the ether.

From the Michelson-Morley experiment and the theory of special relativity, it followed that the ether did not have the simple mechanical properties, which had initially be expected. Some scientists draw – somewhat hastily– the conclusion that an ether did not exist. And in fact, the term “ether” disappeared from many physics text books (although not from all). Thereby, however, the field concept lost its foundation. Previously the field was a special state of the ether, now it went to be a special state of something that does not exist.

However, the logical failure that had arisen was not perceived. A reason may be that Maxwell himself had defined the field as a region of space. It was not noticed that for Maxwell there was no space without ether.

The period of time in which the field had no conceptual basis should not necessarily have been lasted long. At the beginning of the 20th century, it became more and more clear, particularly through the work of Planck about the heat radiation, that the electromagnetic field is a physical system like other systems. But unfortunately, the field concept in the awkward state in which it had gotten shortly after the publication of the special theory of relativity has survived until today.

Besides this complex historical development of the field concept another fact contributes to the confusion:

The term field is not only used as a name for a physical system but also as a mathematical concept. As such, it describes the distribution of values of a physical quantity in space. Thus one speaks of a temperature, a pressure or a density field. Often, the two meanings of the word are not kept apart. Textbooks sometimes mention an “electric field $\boldsymbol{E}$”. But what is meant by that? The physical system “electric field” or the spatial distribution $\boldsymbol{E}(x, y, z)$ of the physical quantity “field strength”.

Disposal:

When introducing the field concept, orient yourself in how you introduce other, material systems. When introducing the ideal gas, one might begin by saying: “An ideal gas is a substance or a system with the following properties…”. Similarly one could introduce the electric field: “An electric field is a system with the following properties…”

Introducing a field as “a region of space with certain properties” would be like introducing a gas, say air, as “a region of space with certain properties”, which is not incorrect; but nobody would do so, with good cause.

[1] *J. C. Maxwell*: A treatise on Electricity and Magnetism, Dover Publications, INC., New York, 1954, p.47

Friedrich Herrmann

3.9 The dipole antenna

Subject:

When introducing electromagnetic waves one often begins with the oscillating circuit. Undamped, damped and driven oscillations are discussed. In order to compensate for the damping, a feedback loop is needed. To reach high frequencies a Hartley type oscillator is used. Next, the coil and the capacitor are reduced to sections of wire, in order to obtain even higher frequencies. Then, the oscillating circuit obtained in this way is bent in such a way that an oscillating dipole is obtained. Thereafter the electric and the magnetic field in the neighborhood of the dipole is discussed, i.e. the „near field". It is then claimed and experimentally demonstrated that an electromagnetic wave is emitted. This wave represents the far field.

Deficiencies:

Several criticisms are expedient.

1) The explanation aims from the beginning at the complicated field of a radiating dipole. Not only do we have to do with an intricate field distribution, but also with the delicate distinction between the near and the far field. Both is not necessary. There are waves that are simpler. In order to discuss wave propagation we propose to begin with simple cases like a plane harmonic wave. There is a wave that is even simpler: the plane square wave.

2) In the traditional introduction of electromagnetic waves, the creation by means of dipole oscillations plays a fundamental role. The idea seems to be that an understanding of the wave is most easily obtained by explaining a procedure for creating the wave. Actually it is much harder to explain this procedure than the wave itself.

This way of arguing is like explaining what an acoustic wave is by beginning with the working principle of a clarinet. A clarinet is a resonator from which a small portion of the energy current flowing back and forth is leaking out and thus is emitted. Just as the clarinet is more difficult to understand than the wave that is produces by it, the dipole oscillator is more difficult to understand than the electromagnetic wave. The dipole antenna is also a resonator from which a small fraction of the energy that is swinging back and forth in the near field is coupled out and emitted.

3) In spite of the great investment that is done when explaining the electromagnetic waves the target is missed. One tries to proceed step by step from the simple oscillating circuit to the supposedly complicated electromagnetic wave. However, the essential step is missing. A gap is remaining. The students learn that due to the electric current that is flowing through the dipole and due to the electric charges accumulating at its ends, the dipole is surrounded by an electric and a magnetic field. Since the explanation is based on the understanding of the oscillating circuit the student may conclude that the phase difference between the electric and the magnetic field is $\pi/2$. However, if this were true, no wave would leave the near field region.

Origin:

After the prediction of electromagnetic waves by Maxwell they had been generated experimentally by Heinrich Hertz. For his experiments Hertz could not take a high frequency generator from the rack. He had to devise an ingenious self-exciting device to create his waves. It is this complicated arrangement that has survived in our textbooks. Being a good theoretician Hertz was also able to calculate the fields of his oscillator, and this calculus also has a prominent place in many textbooks. In Hertz's calculation it can be seen that the phase difference between the electric and the magnetic field deviates from $\pi/2$ already in the near field.

Disposal:

Limit the discussion to waves with a simpler geometry: plane sine or square waves. Prescind from discussing high frequency generators.

The generation of an electromagnetic wave can be explained in the following way: In an extended metal sheet an electric current begins abruptly to flow. Thus a magnetic field will begin to form. This changing magnetic field causes an electric field. The electric field is at all those places where is already the magnetic field. Thus, from the conducting sheet a front is running away that separates that part of the space that is already filled with field and that where is no field yet. When the electric current is switched off again, a second front is moving forward, behind which there is no field. The region between the two fronts represents a plane square wave.

Friedrich Herrmann

3.10 Lenz's law

Subject:

When studying electrodynamics at the secondary school Lenz's law is introduced. According to this law the magnetic field of an induced current opposes the original change in magnetic flux. Lenz's law thus makes a claim about the direction of a vector: the current density vector of the induced current.

Moreover, one shows that the minus sign in Faraday's law of induction

$$U = -\frac{d\Phi}{dt}$$

is a consequence of Lenz's law. The argument goes as follows: Introduce an iron core into a solenoid that is connected to a battery. It is observed that the current intensity decreases for a moment. Since $d\phi/dt$ is positive (so it is argued) and the induced electromotive force (emf) is negative, there must be a minus sign.

Lenz's law is often interpreted in yet another context: Thanks to the minus sign electromagnetic processes obey the law of energy conservation.

Deficiencies:

On the one hand we have to do with a complicated meshwork of statements that got into the school curriculum by tradition, and on the other hand with mistakes in the line of argument.

1. When explaining an algebraic sign in an equation, it should be possible to verify this sign, by measuring the quantities that appear in the equation. The minus sign in the law of induction tells us, that the induced emf has the opposite sign as the time rate of change of the magnetic flux, i.e. $d\phi/dt$. In order to check this claim one must know how to measure both these quantities including the algebraic sign.

Let us imagine that the flux change is as follows: The B vector points in the positive x direction and its absolute value (and thus its x component) increases. What is its algebraic sign? Does this sign remain the same when the coordinate system it rotated by 180°, so that the B vector now points in the negative x direction? Now this flux is surrounded by a circular conductor that is placed in the y-z plane and closed by a voltmeter. How can we read the correct sign from the voltmeter?

Our pupils do not learn how to answer these questions. Therefore the effort in finding out that there should be a minus sign was in vain.

2. In some books we found a mistake in the derivation of the minus sign. This mistake seems to have a life of its own, since although the minus sign can be derived in various different ways, the pattern of arguments, together with the mistake can be found on and off. When the current in the solenoid is switched on, so it is said, $d\phi/dt$ is greater than zero. This conclusion is not correct. The magnetic flux depends depends on the flux density according to

$$\Phi = \int B\, dA$$

Since dA is a vector, ϕ turns out to be positive or negative depending on the orientation of the area of integration.

3. When paying so much attention to the minus sign in one equation, then the plus signs in many other equations should be worth a similar consideration. What does it mean that in Newton's second law, in Ohm's law or in $P = v \cdot F$ there is a plus sign? And which sign do we need to write in Hooke's law? Let us consider a spring that is tended in the vertical direction. We can identify four forces, all related to the spring: the force exerted by the spring on the upper suspension, the force exerted by the spring on the lower suspension, the force exerted by the upper suspension on the spring and the force exerted by the lower suspension on the spring. All of these forces have the same absolute value, two of them are directed upwards, two downwards. So we can take our choice. Although the question of the sign in Hooke's law is not easier to answer than that in the law of induction it is usually dismissed.

4. Why does one insist just in the context of the law of induction that an incorrect sign would mean a violation of the law of conservation of energy? The impression results that induction has something particular to offer in this respect. Actually, there are plenty of other physical laws that would violate the energy conservation law if a sign is inverted: $U = R \cdot I$, $P = v \cdot F$, $F = -D \cdot s$ … Moreover, by changing a sign arbitrarily any other conservation law can be violated.

5. We do not object to honor a scientist by attributing his name to a physical law. But isn't it a little exaggerated to associate the orientation of a current density vector, which anyway follows from Maxwell's equations, to a scientist's name and to give the status of a physical law?

Origin:

Usually a comprehensive theory emerges from predecessors. Unfortunately we sometimes afford the extravagance of teaching not only the last, and in general simplest and clearest version of the theory, but also its predecessors. Lenz's law is only one of many examples.

In 1834, three years after Faraday's discovery of the electromagnetic induction, Lenz had formulated his rule in the following way: "If a constant current flows in the primary circuit A, and if, by the motion of A, or of the secondary circuit B, a current is induced in B, the direction of this induced current will be such that, by its electromagnetic action on A, it tends to oppose the relative motion of the circuits.." [1]

At that time the question for the direction of the induced current was not a trivial one and Lenz's law was a new statement. Only 13 years later Helmholtz made an energy balance and showed that Lenz's rule could also be obtained from energy conservation. Another 25 years later Maxwell published his comprehensive theory of electrodynamics in which Lenz's was absorbed.

The fact that we still today give so much attention to the minus sign in the law of induction is no more than a convention. We must not forget that when teaching, a problem often becomes only a problem if the teacher declares that it is a problem.

Disposal:

Regarding Lenz's law, we refrain from teaching it as a law. Instead we formulate two "hand rules", one for the right and one for the left hand. The right-hand rule is an expression of Maxwell's forth equation, the left-hand rule follows from his third equation.

Right-hand rule: Point the thumb of your right hand in the direction of an electric current. Then the curled fingers point in the direction of the magnetic field vector.

Left-hand rule: Point the thumb of your left hand in the direction of the change dB of the magnetic flux density. Then the curled fingers point in the direction of the electric field vector of the induced electric field.

For the handling of the law of induction there are two possibilities:

Either one explains carefully how one can determine the sign of the various physical quantities, in particular the electric current intensity and the voltage. But that is not all: One also has to explain the convention that associates the algebraic sign of a path integral with that of a surface integral in Stokes' theorem: When the thumb of your right hand points in the direction of surface element of the integral your curled fingers point in the direction of the integration path of the path integral. Regarding the secondary school, we do not recommend to proceed in this way, since it is cumbersome and does not result in any important insight.

So we prefer the second solution: Do not hesitate to formulate the law of induction with only the absolute values of the pertinent quantities.

[1] *J. C. Maxwell*: A Treatise an Electricity and Magnetism, Dover Publications, Inc., New York, 1954, Volume two, No. 542, p. 190

Friedrich Herrmann

3.11 The electromagnet

Subject:

If a core of soft iron is placed inside a coil, the core increases the magnetic field to thousands of times the strength of the field of the coil alone, due to the magnetic permeability μ of the ferromagnetic material.

Deficiencies:

The explanation suggests that the magnetization, and thus the magnetic flux density of the iron increases with the permeability μ. The magnet, so it may seem, has a field that is the stronger, the greater μ is, which is not correct. In order to get an electromagnet from coil, it is sufficient that μ is great compared to 1. It makes almost no difference whether μ is equal to 1000, 10 000 or 100 000.

Origin:

The electromagnet can easily and directly be explained by using Maxwell's equation

$$\oint \mathbf{H} \bullet d\mathbf{r} = \int (\mathbf{j} + \dot{\mathbf{D}}) \bullet d\mathbf{A}$$

which makes a statement about the magnetic field intensity **H**.

Unfortunately, it has become customary to describe magnetic fields mainly by the vector quantity **B**. Sometimes, this restriction is justified by an argument, that should have no place in physics. It is said that **B** *is* the magnetic field. Sometimes it is said that **B** is the fundamental field quantity, whereas **H** is a derived quantity. Here a mistake is made, that we normally do not forgive to our students: A physical quantity is confused with a physical system. Neither **H** nor **B** *is* the field. They cannot be the field, since they are physical quantities, i.e. mathematical objects, invented by man, whereas the field is a physical system, that exists even if no intelligent being is there to observe or to describe it.

The description of the role of the iron core in an electromagnet gets much easier when using **H** instead of **B**.

First, it is easier to define what we understand by a soft-magnetic material: Inside such a material **H** = 0 A/m, whatever the field strength outside of the material is. (This property gets lost when the material gets into saturation. Then the material is not soft-magnetic anymore.)

Second, using **H** we can define a quantity that does not change its value when the iron core is inserted into the coil, i.e. an invariant: The line integral along a curve that loops once around the wires of the coil. If on a fraction of the curve, i.e. inside the iron core, the magnetic field strength is made equal to zero, the contribution to the integral on the remaining part of the curve must increase correspondingly.

The explanation by means of **B** is more complicated and less convincing: It is more difficult to characterize a soft-magnetic material, and there is no invariant when introducing the magnetic core into the solenoid.

Disposal:

When dealing with the magnetism of matter, use **H**, not **B**. When considering **H** as a measure for what we understand by "much field" or "few field", the fact that **H** is zero inside the magnetic core can be formulated as follows: Soft-magnetic materials do not allow the magnetic field to penetrate, just like an electric conductor doesn't allow the electric field to penetrate into it. So it is not difficult to understand the role of the iron core of an electromagnet: Introducing the iron core, while letting the electric current constant, the field is squeezed out of the coil.

Friedrich Herrmann

3.12 Magnetic poles

Subject:

A magnetic pole is either of two regions of a magnet, designated north and south, where the magnetic field is strongest.

Deficiencies:

Magnetization is a vector field that describes the magnetic dipole density within a material. The poles of a magnet are those regions where magnetization lines begin or end. The quantity that allows to describe magnetic poles is the magnetic pole strength or magnetic charge Q_m. The magnetic charge density ρ_m is the source of the magnetic field **H**:

$$\nabla \cdot \mathbf{H} = \frac{\rho_m}{\mu_0}$$

It is the magnetic analogue to the electric charge, or more exactly: to the bound electric charge, as it appears at the surface of a polarized dielectric material. Often, in physics textbooks the magnetic charge is not introduced. Without it, however, it is difficult to describe a permanent magnet quantitatively.

The relation

$\mathbf{F} = Q_m \cdot \mathbf{H}$

which is the magnetic analogue of

$\mathbf{F} = Q \cdot \mathbf{E}$

and which can easily be verified experimentally, cannot be treated. Coulomb's law for magnetic poles is even not mentioned, although it is easier to verify than the corresponding electric law. It is not even possible to define the most fundamental property of a permanent magnet: i.e. that the total magnetic charge of each magnet is zero. Instead, there only remains the rather pale claim that a magnet has (at least) two poles.

For a normal bar magnet the magnetization lines end at the front faces. That means that the magnetic charge is sitting at these surfaces. These, however, are not identical with the regions "where the magnetic field is strongest", since there are field lines that leave the bar magnet at its lateral faces, and as the well-known iron filings pictures show, the field intensity is high at the ends of the lateral faces of the bar magnet. Indeed many students believe that the poles of a bar magnet are also at its sides, and that the pole intensity decreases towards the middle of the magnet. This misconception is further supported by the customary green-red coloring of the side faces of permanent magnets.

Origin:

Formerly, magnetic charge –also called magnetism or pole strength– was introduced in every book about electrodynamics. Maxwell at the very beginning of the second volume of his Treatise introduces Coulomb's law for magnetic poles, (it is the first equation in the book) and he states:

"In every magnet the total quantity of magnetism is zero." [1]

Later the magnetic charge disappeared, due to a misunderstanding. From the fact that no isolated magnetic monopoles, or particles, that carry a net magnetic charge have been found, it was concluded that the quantity, that allows for a description of such a charge does not exist either. However, a physical quantity is not something that "exists" in nature. It is a tool for describing nature, "a free invention of the human mind", as Einstein puts it [2]. Introducing or not a physical quantity is only a question of convenience. And there is no doubt that it is convenient to introduce and to use the quantity "magnetic charge". Without it, we cannot formulate Coulomb's law for magnetic systems, and we can even not pronounce quantitatively the fact that magnets have two poles.

Disposal:

Introduce the extensive physical quantity magnetic charge. Formulate the theorem: "The total magnetic charge of a magnet is zero."

[1] *J. C. Maxwell*: A Treatise on Electricity and Magnetism, Dover Publications Inc., New York, 1954, p.4.

[2] *A. Einstein*: Mein Weltbild, Ullstein Taschenbücher-Verlag, 1957, p. 115.

Friedrich Herrmann

3.13 The field of permanent magnets

Subject:

In physics textbooks for the secondary school field line pictures of bar magnets are shown. Examples can be seen in Fig. 1.

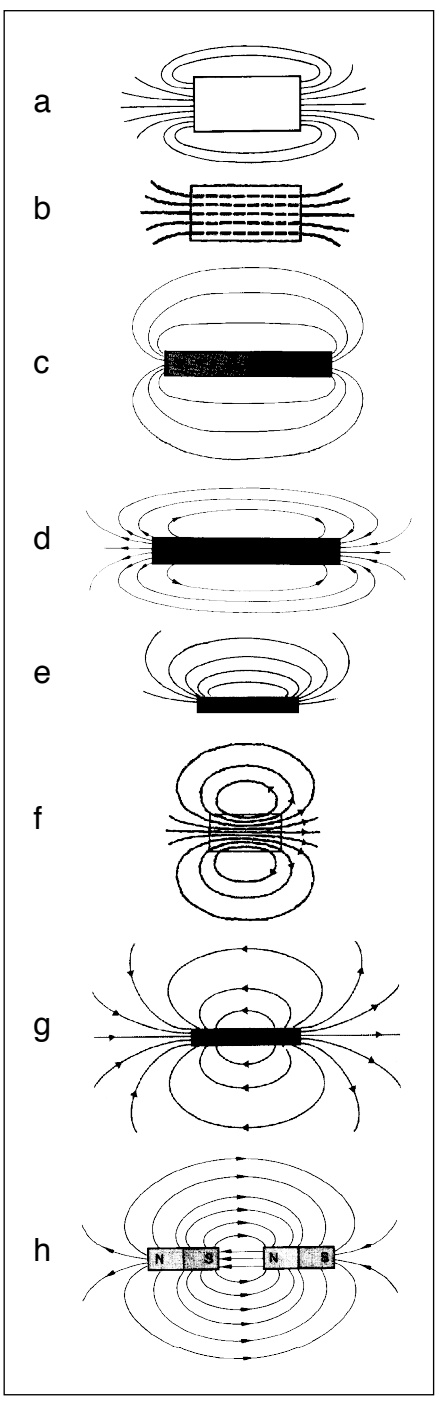


Fig. 1. Magnetic field lines of bar magnets, as shown in secondary school books

Deficiencies:

In all the text books that I have consulted the corresponding pictures are incorrect. (I have examined ten school books, most of them German, some American and one Italian). The errors can be seen when comparing the school book pictures of Fig. 1 with the correct drawing of Fig. 2a (which was taken from Sommerfeld [1]). The difference between Fig. 2a and the school book pictures do not allow the excuse that different pole distributions have been presupposed, since there are no pole distributions that correspond to fields as shown in Fig. 1.

In the various text books different errors can be detected.

1. Field lines exit or enter the magnet only at the end faces, Figures 1a, 1b and 1c. Actually they also do so at the lateral faces.

2. Field lines exit the magnet only perpendicularly to the surface, Figures 1a and 1b. Actually they are perpendicular only at the center of the end faces.

3. Those field lines, which enter into or exit from the magnet at the lateral faces have an incorrect orientation, Figures 1d, 1e, 1f, 1g and 1h.

Similar errors can be found in the pictures of the magnetic field of the horseshoe magnets and of the Earth. Often, a school book also shows a photograph where the field lines are made visible by means of iron filings. In such books one may see on two figures that are side by side the discrepancy between the real field and what the field line picture pretends.

One of the pictures that I have found, displays even more crude errors: Field lines do not only begin at the north pole but also end there, and the same happens at the south pole.

Origin:

A graphical representation of a physical phenomenon must not necessarily be precise in every respect. It must show and may emphasize the essential. Unimportant details may be omitted for the sake of clearness. In the present case, however, no simplification was made, but a message was conveyed that is not correct. It is not true that the students do not perceive the incorrect claims. One can easily convince oneself that they keep them in mind. Indeed, many students believe, that the field lines depart perpendicularly from the end faces of a bar magnet. When asking a student to draw a field line picture of a bar magnet, almost always an incorrect picture or sketch is drawn. Obviously, the students draw something which they have memorized.

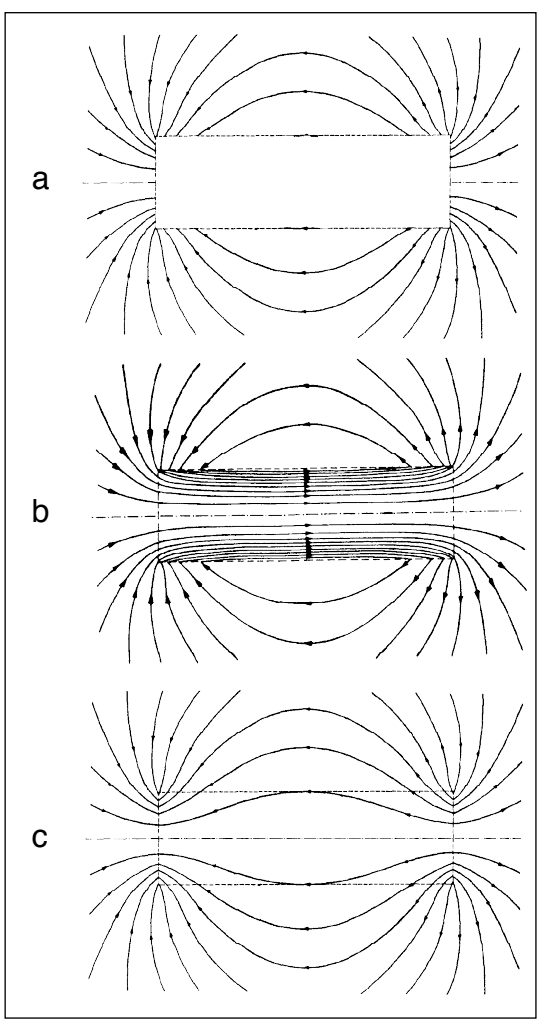


Fig. 2. Bar magnet. (a) **H** and **B** field lines outside the magnet; (b) **B** field lines; (c) **H** field lines

When asking for an explanation or justification of the direction of the field lines, their reaction is perplexity. Actually, the incorrect pictures have a certain plausibility.

The incorrect direction of the field lines that exit or enter at the lateral faces might be justified in the following way: The students know that the **B** field is divergence-free. The **B** field lines have no beginning and no end. Thus, they can be completed or continued at the inside of the magnet. Now, they make the incorrect hypothesis that the field lines have no kink when crossing the surface of the magnet. And indeed, in one of the text books, the field lines had been drawn in this way, Fig. 1f. The field lines would be even more “smooth”, if they left the magnet only at the end faces, as in Fig. 1b. (The figures 1b and 1f are taken from the same book, but they do not agree with one another.) The correct shape of the lines is shown in figure 2b. Notice the pronounced kinks of the lines at the lateral surfaces.

Those who let the lines enter and exit only at the end faces may believe that inside they are identical with the magnetization lines, what is not true. (The magnetization lines form a homogeneous field.)

Those who let the lines enter and exit perpendicularly to the surfaces may believe, that a rule applies which we know from the electric field lines at the surface of an electric conductor.

We were surprised to find that all the University text books that we had consulted show the correct pictures, whereas all the school books show incorrect pictures. Apparently there exists something like a “school physics” that has its own life, independent from “University physics”. It also shows that a “new” book in general is not really new. It contains the old errors in a new packing.

Disposal:

Draw correct field lines. A help may be: Do not draw the **B**, but the **H** lines, Fig. 2c. The magnetic poles are the sources of the **H** field lines. One may imagine that the end faces are not magnetic poles but carry electric charge. The problem of drawing the **H** field lines is the same as that of drawing the electric field lines for the charged end faces.

[1] *A. Sommerfeld*: Vorlesungen über Theoretische Physik, Band III, Elektrodynamik. – Akademische Verlagsgesellschaft,
Leipzig 1964. – S. 78

Friedrich Herrmann

3.14 Equipotential surfaces

Subject:

To represent an electric, a magnetic or a gravitational field graphically commonly field line pictures are drawn. In the case of the electric and the gravitational field sometimes equipotential surfaces are also represented.

A field line picture expresses two aspects of the field:

1. It shows the direction of the field strength vector in each point of the field. The field strength vectors point in the direction of the tangents of the field lines.

2. It tells us where the “sources” of the field are located. These are the places where the field lines begin or end.

Sometimes it is said that a field line picture allows to read the magnitudes of the field strength vectors. Actually this is true only in special cases [1,2].

Deficiencies:

A graphical representation of a field allows us to grasp at a single glance, what would be complicated to express in words. (“A picture is worth a thousand words.”) However, although there are several possibilities to graphically represent a field mostly a single method is used: the field line picture. We are so accustomed to this representation that it hardly comes to our mind, that there are alternatives. One such alternative are the surfaces that are orthogonal to the *field lines*, the *field surfaces*.

Fields are often introduced as rather abstract entities. Therefore for many students the field lines are the straw which they will catch. And the result is often that they identify the field lines with the field.

Origin:

For Maxwell it was natural to represent all fields by both the field lines (“lines of force”) and the field surfaces (“equipotential surfaces”), Fig. 1. This was a method to realize a suggestive picture of an invisible object. At the turn of the century serious doubts about the existence of an aether came up, and the aether was banned from physics. As a consequence, the field degenerated into an abstract entity, hardly more than a mathematical concept for calculating forces. From now on field lines were no more than auxiliary lines that represented the direction of the force on a test particle. The orthogonal surfaces survived only in the form of equipotential surfaces in the special case that the fields were conservative. Since a potential can only be defined for a conservative field, the opinion was now, that drawing the orthogonal surfaces makes sense only in such fields. Apparently, it was not noticed that the only problem was the name. Actually orthogonal surfaces can also be drawn for nonconservative fields. Then they are not equipotential but they are just as useful as the equipotential surfaces in conservative fields. Actually, they become particularly interesting for nonconservative fields, since they indicate clearly where the curl of a field is located.

Disposal:

In the following we call *flux sources* those places within a field where the divergence is different from zero. We call *circulation sources* the places where the curl of a field is different from zero.

Just as flux sources are places where field lines begin or end, circulation sources are places where field surfaces terminate. In a field line picture the flux sources are easily seen, whereas the field surfaces indicate clearly the circulation sources. That is why one best represents both in each field picture: field lines and field surfaces (in a two-dimensional plot the field surfaces also appear as lines).

Consider as an example electric fields. The flux sources are electric charges, the circulation sources are places where the magnetic flux is changing with time.

Fig. 2 shows two linear charges (thin charged wires, perpendicular to the drawing plane) and three thin “linear” coils whose magnetic flux is changing with time. The coils are also perpendicular to the drawing plane, and therefore appear as points.

The Figure can also be interpreted as a magnetic field. Then the flux sources are magnetic line charges (linear magnetic poles), the circulation sources are electric currents.

[1] *A. Wolf, S. J. van Hook, E. R. Weeks*: Electric field line diagrams don't work, Am. J. Phys. **64** (1996), p. 714 - 724.

[2] *F. Herrmann, H. Hauptmann, M. Suleder*: Representations of Electric and Magnetic Fields, Am. J. Phys. **68**, p. 171.

Friedrich Herrmann

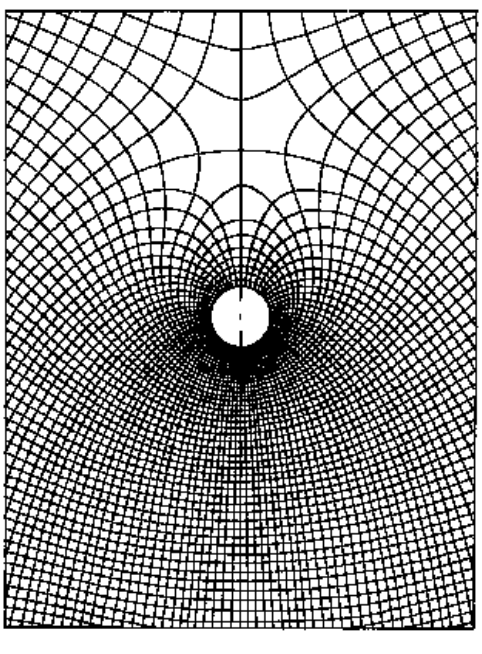


Fig. 1. Superposition of the magnetic field of an electric conductor (perpendicular to the drawing plane) and a homogeneous magnetic field from Maxwell’s *Treatise on Electricity and Magnetism*

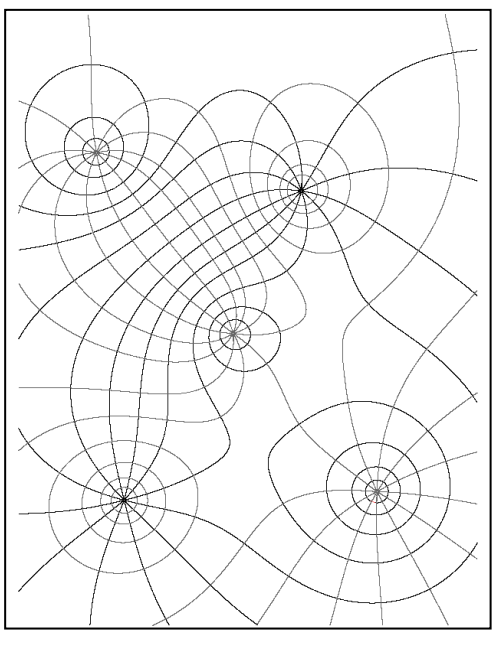


Fig. 2. Field of two flux sources and three circulation sources (field lines: black; field surfaces: grey)

3.15 Inductivity

Subject:

In school books the inductivity L is usually introduced by means of the law of induction, namely as the factor of proportionality between the induced "electromotive force" U_{ind} and the time rate of change of the electric current which is the cause of the induction:

$$U_{ind} = -L \cdot \frac{dI}{dt} \qquad (1)$$

Deficiencies:

Electric engineers know three passive linear electric components: the resistor, the capacitor and the coil. (The mechanical analogues are the vibration damper which obeys Stokes' law, the mass point and the spring which obeys Hooke's law.) For each of these component a linear relation holds:

resistor: $U = R \cdot I$

capacitor: $Q = C \cdot U$ (2)

coil: $n\Phi = L \cdot I$ (3)

R, C and L depend on the geometric dimensions and the material of the components. Whereas in the resistor energy is dissipated, the capacitor and the coil can store energy. For circuits which contain, apart from a power supply, only components of these three kinds, an internal symmetry holds: If the circuit is replaced with another one according to certain translation rules, the new network is described by equations which have the same mathematical structure as those of the initial network. A well-known example is the RC circuit which transforms into an RL circuit. The most important translation rules can be read from the following table:

U (voltage)	⇔	I (electric current intensity)
Q (electric charge)	⇔	$n\Phi$ (magnetic flux)
C (capacity)	⇔	L (inductivity)
R (resistance)	⇔	$1/R = G$ (conductance)
junction	⇔	loop
series circuit	⇔	parallel circuit
voltage stabilized power supply	⇔	current stabilized power supply

The double arrow is to read as follows: U has to be replaced with I und I with U, Q with $n\Phi$ and $n\Phi$ with Q, etc.. The appearance of the number of turns n which somewhat disturbs the aesthetics is due to the fact that by the flux through a coil we mean the product of the flux density B and the cross sectional area of the coil. It would be more logical to use the quantity $\Phi' = n\Phi$ since the effective surface crossed by the field lines is n times the cross sectional area of the coil.

When defining the inductivity by equation (1) this beautiful symmetry is ignored. The analogy between the capacitor and the coil is less evident.

The inconvenience of introducing L via equation (1) can also be seen when trying to introduce the capacity in a corresponding manner, i.e. by means of that equation which is analogue to equation (1):

$$I = C \cdot \frac{dU}{dt}$$

The equation describes the process of charging or discharging of a capacitor. If we use it for the introduction of C, we suggest that the capacity is a quantity that is interesting only for such processes. The introduction by means of equation (2) is more general. Here C appears as a measure of the effort we have to do in order to store a given electric charge: Do we need a higher or a lower voltage?

The same is true for the inductivity. When introducing it by means of equation (1) we suggest that L is a quantity that is only important in the context of electromagnetic induction. Equation (3), which is used for the introduction of L in some university text books, tells us on the contrary something about L without referring to induction. It appears as a measure for the effort we have to do in order to establish a given magnetic flux in a coil. Do we need a stronger or a weaker electric current?

Origin:

We can introduce the inductivity, just as the capacity, the resistance and many other physical quantities by any equation in which the quantity appears. Apart from equations (1) and (3) the inductivity is sometimes introduced by a third equation:

$$E = \frac{L}{2} \cdot I^2 \qquad (4)$$

This equation gives the energy stored in the field of a coil. These three possibilities of the introduction of L (by equations (1), (3) or (4)) coexist since the appearance of L in physics. Here again it can be noticed that in some respects school and university physics have a rather independent existence.

Disposal:

We introduce the magnetic flux as $B \cdot A$. We show experimentally that the flux density in a coil is proportional to the intensity of the electric current in the coil. It follows that also the flux $n\Phi$ is proportional to the current intensity:

$n\Phi \sim I$.

The factor of proportionality is called inductivity:

$L := n\Phi/I$.

In order to get equation (1) we have to combine with the law of induction

$U_{ind} = nd\Phi/dt$.

Friedrich Herrmann

3.16 Magnetic poles of a solenoid

Subject:

The magnetic field of a solenoid is in the space outside of the solenoid identical with that of a bar magnet with the same geometrical dimensions. When introducing the solenoid and its field at school this fact is usually mentioned. In addition it is often said that the solenoid has a north and a south pole at its ends. The corresponding statement is sometimes made for a current loop. Here, the poles are on both sides of a circular surface defined by the current loop.

Deficiencies:

To find out where are the magnetic poles of any arrangement, one best looks for the magnetization **J**. Magnetization is a vector quantity that describes the magnetic state of matter. It tells us which is the magnetic dipole moment of each volume element. Fig. 1 shows the magnetization of a magnetic disk whose poles are both at the lower side. It also shows the magnetization lines. These lines always begin south poles and end on north poles. Neither the solenoid nor the current loop have a magnetization. Thus, they do not have magnetic poles.

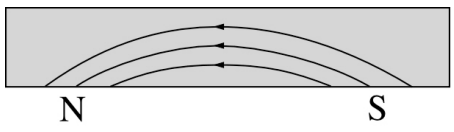


Fig. 1.

To localize the magnetic poles one can also look at the **H** field line picture. **H** lines begin at north poles and end at south poles. For a solenoid and a current loop the divergence of **H** is zero everywhere; thus we can again conclude that there are no poles.

It is true that the magnetic fields of a solenoid and a bar magnet are similar, and it is worth mentioning this fact. However, when attributing poles to the solenoid the students will get an erroneous idea about what is a magnetic pole. They will not understand a fundamental difference between an empty solenoid and an electromagnet.

And finally: If a solenoid and a current loop would have poles, then one should expect that any other current distribution also must have poles. Where are the poles of a current-carrying wire? Usually we emphasize that there are no poles.

Origin:

The comparison between the fields of permanent magnets and current distributions is a standard subject of university physics. It is an important subject since students learn to distinguish between the divergence and the curl operator. Which distribution of divergences leads to the same field as a given distribution of curl?

Apparently school physics has borrowed from this subject, but the authors are not aware of the errors they make.

The idea that currents cause magnetic poles is further kept alive by the fact that in Geography it is common to speak of the magnetic north and south pole of the earth. These are no poles in the sense of physics for two reasons: First, they are caused by electric currents. Thus the magnetic field of the earth is divergence-free. And second, according to the geographical definition they are points at the earth's surface (those point where the horizontal component of the magnetic field strength is zero). Even if the cause of the magnetic field of the earth were of ferromagnetic origin, the poles would not be points, but extended regions within the earth.

Often it is said that the earth is a magnet. This idea goes back to Gilbert, who found that the origin of the magnetism of the earth resides in the earth's interior, and not in the sky. He conjectured that there is a magnet within the earth. His opus "De Magnete…" appeared in 1600, i.e. 220 years before Oerstedt's discovery of the relation between electric currents and magnetic fields, and a long time before it was known that the interior of the earth is so hot that no magnetized material can exist. If we look for a comparison, instead of saying the earth is a magnet it would be more convenient to compare the earth with a current carrying ravel.

Disposal:

Do not say a solenoid or a current loop has poles. When discussing the electromagnet, instead of saying the electromagnet has poles it is better to say that its iron core has poles.

Friedrich Herrmann

3.17 Leakage field of the transformer

Subject:

"…, the magnetic flux Φ should be completely confined to the interior of the iron core, i.e. run through both windings with the same intensity (no leakage flux)." [1]
"When measuring the secondary voltage more precisely it turns out to be smaller, than what would be expected from the calculus: This fact is due on the one hand to Joule losses,…. The second cause is that, due to leakage, only a part of the induction flux in the primary winding crosses the secondary winding." [2]

Deficiencies:

Students learn that leakage fields or stray fields are something that should be avoided. In principle, they are not necessary, and no fundamental physical principle is violated if we imagine a physical world without them. It is similar to mechanical friction. Also friction often appears only as a nuisance, which one tries to prevent. A somewhat rougher analogy is a leak in a garden hose. Admit the hose has some perforations or the fittings are not watertight. These leaks can, in principle, be completely tamped. And indeed, this can also be said for certain "stray fields". A metallic shielding prevents the leakage of an electric field, a Mu-metal shielding encloses a magnetic field or holds it off from some other device.

Now, the same name "leakage field" or "stray field" is also used in cases where where the working principle of an apparatus depends on this field. Among several examples there is the transformer, which we will discuss in the following.

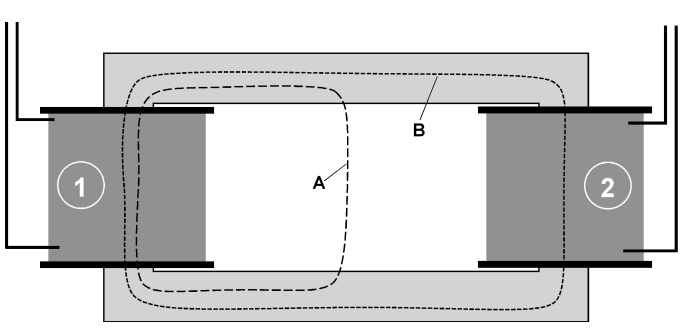


Fig. 1. The magnetic field strength *H* is different from zero only on that part of the integration path which runs outside of the iron core.

We consider a transformer in its most simple form: an iron core that forms a closed rectangle, with the two windings sitting on the opposite shorter sides, Fig. 1. We make the conventional assumptions:

– the ohmic resistance of the coils is small compared with the respective inductive resistances;
– the load resistance is small compared with the inductive resistance of the secondary coil;
– the load resistance is great compared with the ohmic resistances of either coil;
– the permeability μ of the core material is much greater than one.

We now apply Ampere's law, and first integrate along path A:

$$\oint_A \vec{H}\, d\vec{r} = n_1 I_1$$

The value of the integral is equal to the total current $n_1\ I_1$, that is looped by the integration path. (n is the number of windings, the indices refer to the primary and secondary windings, respectively.) Now, the magnetic field strength H inside the iron core is smaller than outside by the factor μ. Since typical values of μ are greater than 1000, the contribution of the path inside the iron to the integral is negligible. Thus, only the "leakage field" contributes to the value of the integral.

We now consider integration path B. It loops through both of the coils. Since it passes on its entire length within the iron core the integral is equal to zero:

$$\oint_B \vec{H}\, d\vec{r} = n_1 I_1 - n_2 I_2 = 0$$

We thus get the well-known relation:

$n_1\ I_1 = n_2\ I_2$.

We see that this relation could not hold if there were no "leakage field".

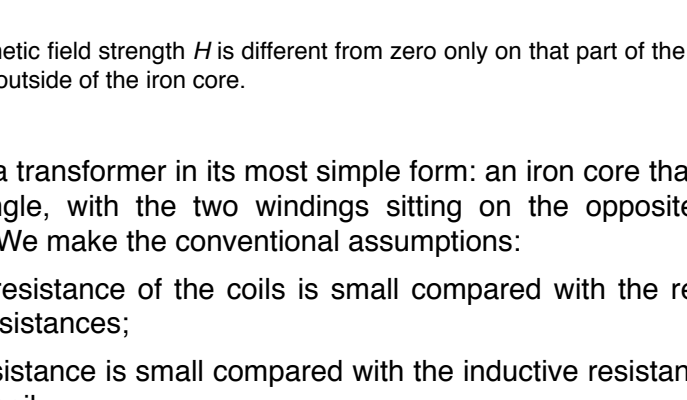


Fig. 2. The fields between the legs of a transformer. The energy flow is from left to right.

The importance of the denigrated field can be seen in yet another way. Fig. 2 shows schematically the H field lines as well as the electric field lines. The change of the magnetic flux within the iron core is the cause of an electric eddy field, whose field lines loop around the legs of the iron core. In addition, the figure shows the Poynting vector:

$$\vec{S} = \vec{E} \times \vec{H}\,, \qquad (1)$$

i.e. the energy flow density within the field. It is seen that the energy gets from the primary to the secondary circuit through the field.

The situation is analogous to that of the energy transport by means of an electric cable. The only difference is that the electric and the magnetic fields are interchanged, Fig. 3. Since the conductors have different electric potentials, the electric field lines run from one conductor to the other, and since an electric current is flowing in the conductors, they are surrounded by a magnetic eddy field. The energy flow distribution is the same as in the transformer.

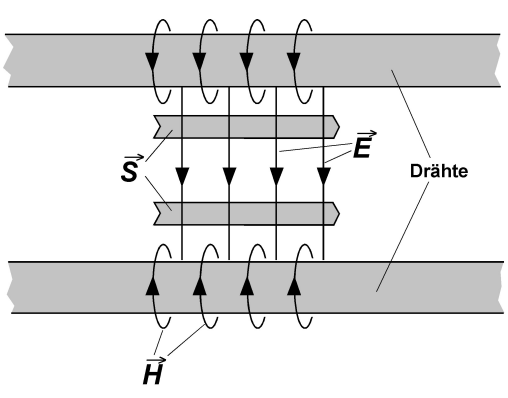


Fig. 3. The fields between the two conductors of an electric cable. The energy flow is from left to right.

The stray field of the transformer is not more responsible for the losses of the transformer than the electric field between the conductors of a cable are responsible for the losses of the cable. In both cases the efficiency is essentially limited by the dissipation within the "conductors". In the transformer we have energy dissipation in the coils and in the iron core due to the steady change of the magnetization. A measure of this dissipation is the magnetic field strength within the iron core. Ideally it should be zero, just as the electric field strength within the electric conductors of the cable should be zero. Since the dissipation in an iron core is rather high, in technical transformers the distance between the primary and the secondary coil is made as short as possible.

Origin:

In the common discussion of the working principle of the transformer one does not argue with the magnetic field strength H, but only with the flux density B. Since B is much greater inside the iron core than outside the impression results that the field outside does not play an essential role. This is only one of many examples which show that the exclusive use of B for the description of magnetic phenomena causes misconceptions. Another cause may be that one avoids the discussion of the local energy balance.

Disposal:

1. Do not throw all the "stray fields" into the same pot. Since the terms "stray field" and "leakage field" have a negative connotation, it would be better not to employ these words to the field between the legs of a transformer.

2. Do not limit the discussion of the magnetic field of the transformer to B. Discuss also H.

3. Ask as often as possible the questions: Where is the energy? Which way does the energy go?

[1] *Gerthsen:* Physik, 21. Auflage, Springer-Verlag Berlin 2002, S. 414.

[2] Handbuch der experimentellen Schulphysik, Elektrizitätslehre III, Aulis Verlag Deubner & Co KG Köln 1965, S. 70.

Friedrich Herrmann

3.18 Force fields

Subject:

"Experience tells us that the […] forces, that act on a mass point m can depend on the position vector $\boldsymbol{r}$ of the mass point, and/or its velocity $\dot{\boldsymbol{r}}$, and also on time. Thus, in general, it will be a force $\boldsymbol{F} = \boldsymbol{F}(\boldsymbol{r},\dot{\boldsymbol{r}},t)$." [1]

"In physics a force field is a vector field that describes a non-contact force acting on a particle at various positions in space. Specifically, a force field is a vector field $\boldsymbol{F} = \boldsymbol{F}(\boldsymbol{x})$, where $\boldsymbol{F}$ is the force that a particle would feel if it were at the point $\boldsymbol{x}$." [2]

"Restating mathematically the definition of energy (via the definition of work), a potential scalar field $U(\boldsymbol{r})$ is defined as that field whose gradient is equal and opposite to the force produced at every point…." [3]

Deficiencies:

Each physicist has experienced a lecture about analytical mechanics. There he or she learned, among other things, what our citations are telling: A force depends on position, and sometimes also on velocity and time.

Whenever we indicate the value of a physical quantity, it must be clear to what entity this value refers. There are the local quantities, whose values refer to a point, such as temperature, pressure or electric field strength. Other quantities refer to a surface area. All the currents and fluxes belong to this class: electric current, power (= energy current), magnetic flux, and force (momentum current). The values of another class of quantities refer to a region of space. These are the extensive quantities mass, energy, electric charge, entropy etc. There are quantities for which the assignment is more complicated, as for instance electric resistance, capacitance etc.

Our actual subject is force. Force $\boldsymbol{F}$ is related to mechanical stress $\boldsymbol{\sigma}$ by

$$\boldsymbol{F} = \iint_S \boldsymbol{\sigma}\, \mathrm{d}\boldsymbol{A}\,. \qquad (1)$$

Mechanical stress $\boldsymbol{\sigma}$ is a local tensor quantity, the surface element $\mathrm{d}\boldsymbol{A}$ is a vector quantity. Thus, the force in equation (1) refers to surface area S.

In the case of a rod that is under uniform compressive or tensile stress in the direction of its length equation (1) simplifies to

$$|\boldsymbol{F}| = \sigma \cdot A\,.$$

where σ is the stress component in the direction of the rod and A is its cross sectional area. In liquids and gases at rest the components of the stress tensor corresponding to the three directions in space are all equal and the tensor is always diagonal. In this case the stress is called hydrostatic pressure p:

$$\boldsymbol{F} = p \cdot \mathbf{A}$$

Whatever the orientation of the reference surface, the force has the same direction as the surface area vector.

Forces that are transmitted by electromagnetic fields can also be calculated by equation (1). In this case, $\boldsymbol{\sigma}$ is the Maxwell stress tensor. If the surface S is chosen in such a way that it encloses the whole of a body, one gets "the force that acts on the body".

Our considerations show that when specifying a force, we have to indicate the surface to which it refers. However, this statement is in contradiction to our citations, which make sense only if force is a local quantity, i.e. if its values refer to a point.

Our citations belong to analytical mechanics. In this context, force can indeed be introduced as a local quantity as long as one operates with the model of point masses and point charges. Then both quantities, mass and force, refer to a point. On the contrary, in continuum mechanics, mass is an extensive quantity and thus refers to a region of space. As a consequence force refers to a surface area.

Point mechanics is treated at the University so extensively that it is easily forgotten that one has to do with a model, that is indeed very useful, but at the same time conceptually somewhat strange. Why strange? Some quantities that are known in the "normal", i.e. continuum mechanics, become infinite, or better, they do not exist: densities, current densities and mechanical stress.

Since in point mechanics a force refers to a point (and not to an area), it is possible to attribute a force to every point in space by means of a point-like test body. In this way one gets the function $\boldsymbol{F} = \boldsymbol{F}(\boldsymbol{r})$, called "force field".

We know that force fields play an important role in Hamilton and Lagrange theory, and that many real systems can be described in a very good approximation as systems of point masses. However, we thereby may lose sight of the fact that statements as those of our citations are not generally valid but are tailored for point mechanics.

Origin:

Newton attributed a force to a body and not to a point. Since at that time the concept of field did not yet exist, Newton's forces could not refer to a reference surface. So he attributed the force to the body instead of its boundary surface.

Point mechanics which attained its full blossom with Lagrange, Hamilton and Jacobi, and later served as a master for quantum mechanics, was so naturally taught as the genuine mechanics, that some awkward consequences of the pointlikeness of the bodies are easily overseen.

Disposal:

When introducing a new physical quantity make clear to what geometrical entity its values refer. The necessity of such a procedure can be shown easily: Establish with your students or pupils a list with all the physical quantities they know. Then ask them to tell about each of these quantities to what geometrical entity they refer. Probably, you will experience an unpleasant surprise.

Instead of operating with force fields, consider the corresponding field strengths (electric, magnetic or gravitational). These are indeed local quantities.

[1] *C. Schaefer, M. Päsler*: Einführung in die Theoretische Physik, Verlag Walter de Gruyter & Co 1970, S. 92

[2] Wikipedia, Search term *Force field*

[3] Wikipedia, Search term *Force*

Friedrich Herrmann

3.19 Two phenomena of electromagnetic induction

Subject:

When introducing electromagnetic induction often two realizations of induction experiments are distinguished.

In one of them an electric conductor is moved through a uniform magnetic field that is constant in time, Fig. 1a. The interpretation of the experiment is as follows: A Lorentz force acts on the charge carriers and displaces them until the electrostatic force which results from the displacement equilibrates the Lorentz force. Between the ends of the conductor an electric potential difference has built up that can be measured with a voltmeter.

In the second induction experiment the electric conductor remains at rest, and the magnetic field strength is changed, by moving the magnet, Fig. 1b. Again the voltmeter deviates. This experiment cannot be interpreted with a Lorentz force. It seems to be based on another physical effect.

However, both results can be summarized in one equation, the law of electromagnetic induction:

$U_{\mathrm{ind}} = -\,\mathrm{d}\phi/\mathrm{d}t$

This procedure is commented in the following way: “Surprisingly, two different physical causes of the electromagnetic induction can be summarized in a single equation.“

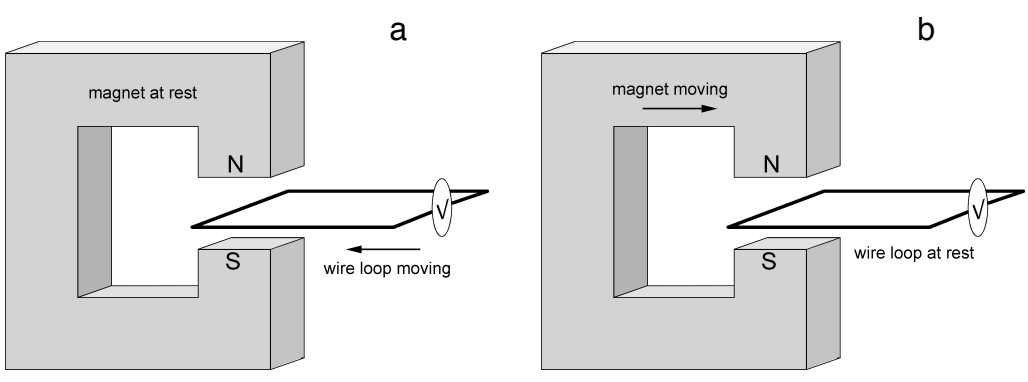


Abb. 1. Electromagnetic induction in two different reference frames. (a) Magnet at rest, wire loop moving; (b) mangnet moving, wire loop at rest.

Deficiencies:

The same experiment is described in two different reference frames.

We begin by considering the second experiment, Fig. 1b. The wire loop is at rest and the magnet is moving. For the interpretation the second Maxwell equation is needed:

$$\mathrm{rot}\,\vec{E} = -\dot{\vec{B}}\,.$$

The magnetic flux density inside of the wire loop is changing. Thereby a non-conservative electric field ($\mathrm{rot}\,\vec{E} \neq 0$) is created. By means of the integral form of Maxwell’s second equation

$$\oint \vec{E}\,d\vec{r} = -\dot{\Phi}$$

this fact can be expressed as follows: The magnetic flux that traverses the wire loop is changing. Thereby within the conductor a emf is created.

Now the experiment of Fig. 1a: A Lorentz force acts on the charge carriers. This is equilibrated by an electrostatic force. There is a conservative electric field ($\mathrm{rot}\,\vec{E} = 0$). The magnetic flux density $\vec{B}$ does not depend on time. In order to conciliate the version of the experiment with Maxwell’s second equation one often used a somewhat inelegant mathematical trick. When calculating the magnetic flux as the surface integral of the flux density one admits that the integration surface changes with time. In a strict sense this corresponds to a hidden change of the reference frame.

We see that one and the same experiment was described in two different reference frames. When passing from one frame to the other field strengths transform, i. e. change their values. Only in this way is it possible that in one case we have a conservative electric field and in the other a non-conservative.

To better understand the consequences of a change of the frame of reference, let us consider an experiment that is even more simple:

A single magnetic north pole P (the end of a long permanent magnet) is moving relative to a small body Q that is positively charged, Fig. 2. The movement is perpendicular to the straight line connecting P and Q. We describe what happens in two reference frames: the frame in which P is at rest (upper figures) and that in which Q is at rest (lower figures).

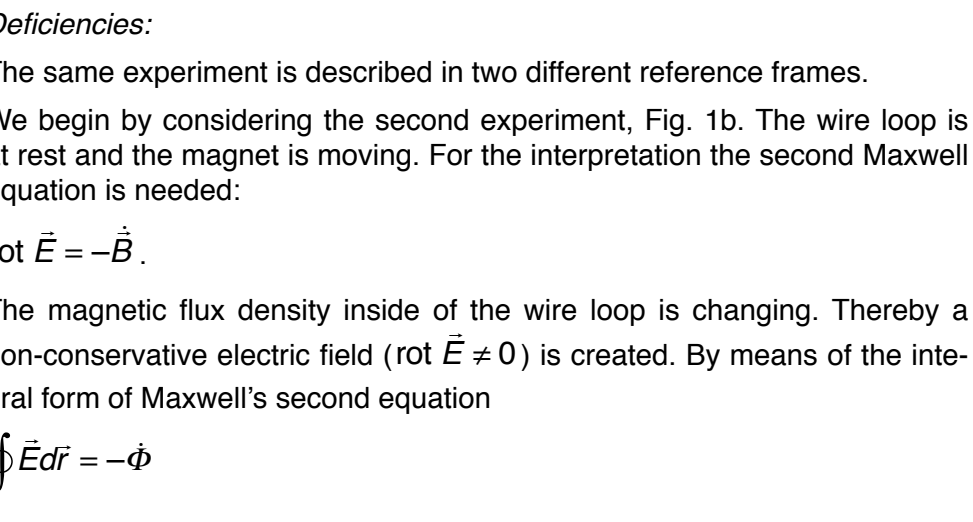


Abb. 2. A small body Q that is electrically charged moves with respect to a magnetic pole P. The process is represented in two different reference frames. (a) The magnet is at rest, the charged body moves in the direction perpendicular to the drawing plane. (b) Body Q is at rest, the magnetic pole moves out of the drawing plane. The lines in the upper figures (a) are magnetic, in the lower figure electric field lines. The two drawings at the left represent the contributions of P and Q separately. The figures at the right show the resulting field.

1. Reference frame of P

The electric charge that moves with body Q (Fig. 2a, movement into the plane of drawing) represents an electric current. This current is surrounded by a magnetic field. (Maxwell’s *first* equation is responsible for this field.) P “feels” this field and experiences a force in the upwards direction. Meanwhile a Lorentz force that is oriented downwards acts on Q. Thus, the interaction between P and Q is mediated by a magnetic field. In the left part of Figure 2a the contributions or P and Q to the magnetic field are represented separately. In this way the force on P in the field of Q and the force of Q in the field of P can be read. These forces can also be read from the right part of the figure, in which the resulting field strengths are represented. The field lines are denser above Q than below. Since there is compressional stress in the direction perpendicular to the field lines, Q is pushed by the field downwards. Moreover, the field lines are denser above P than below. Since there is tensional stress in the direction of the field lines, P is pulled upwards by the field.

2. Reference frame of Q

The moving magnetic pole P represents a magnetic „displacement current“ (Fig. 2b, movement out of the drawing plane). This is surrounded by an electric field. (Maxwell’s *second* equation is responsible for this field.) Q feels this electric field and experiences a force in the downward direction. Meanwhile a force that is oriented upwards acts on P: The electric analogue to the magnetic Lorentz force (magnetic current within an electric field). Thus, the interaction between P and Q is mediated by an electric field. Here too, the forces can be read from left as well as from the right figure.

The example shows, that in electrodynamics a change of the reference frame can require that a phenomenon has to be described sometimes by means of Maxwell’s first equation and sometimes by Maxwell’s second equation, and that one and the same interaction is mediated sometimes by an electric and sometimes by a magnetic field.

Origin:

We usually discuss reference frame changes only in mechanics, and are not trained to at reference frame effects in other fields of physics, such as electrodynamics or thermodynamics.

Disposal:

We know from mechanics that a change of the reference frame brings complications, and that an inappropriately chosen of the reference frame can make the description of a phenomenon cumbersome. If the change of reference frames is not the actual teaching objective, we recommend to elude the subject.

Friedrich Herrmann

3.20 Conservative vector fields

Subject:

Physics students learn that for an induced electric field the electric potential is not defined: “The existence of rotational electric fields shows that not every field has an electric potential […]. Such fields have closed field lines. An electric charge can gain any amount of energy when moving on a closed path.”

Deficiencies:

We limit ourselves to consider electric fields. Similar arguments hold for magnetic fields and also for velocity fields of flowing liquids.

Among the electric fields there are two classes with particular properties: Conservative fields and rotational fields. A conservative electric field is a field for which

$$\nabla \times \vec{E} = 0$$

everywhere. That means that the field must have sources somewhere, i.e. $\nabla \cdot \vec{E}$ cannot be zero everywhere. Otherwise there would not be a field at all.

The places where $\nabla \cdot \vec{E} \neq 0$ are sometimes called *flux sources*.

A pure rotational field is a field for which

$$\nabla \cdot \vec{E} = 0$$

everywhere and

$$\nabla \times \vec{E} \neq 0$$

somewhere.

We call the places where $\nabla \times \vec{E} \neq 0$ *circulation sources*.

In general, a field will have both kinds of sources and thus will not belong neither to the one nor to the other category.

Nevertheless, these concepts play an important role in the teaching of electrodynamics. The reason is that one often imagines that there is nothing else in the world than an electric dipole, a plate capacitor or a current-carrying solenoid. About their field simple statements can be made as for instance the following: The electric field of an electric dipole is a conservative field, or the electric field around a solenoid whose electric current is changing in time is a rotational field.

The simplicity of this classification sometimes gives rise to a conclusion that overshoots the target. An example is our citation: An induced electric field has no electric potential.

In order to see the problem we first have to obtain clarity about how we want to employ the word “field”.

Sometimes we speak of the electric field of a point charge, of a dipole or a capacitor (or of the magnetic field of a solenoid, a current loop or a bar magnet). When doing so we imagine that there is nothing else in the world than this point charge, dipole etc.

In other occasions we speak of the field in a given domain of space and it may be that the sources of this field are not our primary concern.

Statements as that of our citation refer to the first of these situations. They are global and general statements and they refer to systems of infinite extension. They are reasonable and useful when the intention is to get certain general insights about electrodynamics, but sometimes they are inappropriate. When dealing with a practical problem we are not interested in a statement about the world at large, but only about a given region of space. So a practical question may be: Does the region of space that we consider contain any flux or ciculation sources? If there are no circulation sources within this region, we can define a potential. If the curl of the field is non-zero only at certain places within the considered space then we can cut out a simply-connected region that does not contain curl sources and define a potential for this region. Whether there are circulation sources outside of our region does not have any importance for our decision in favor or against a potential. If we took the statement of our citation at face value, one could never employ the useful tool “potential”. It would be forbidden to say that the neutral conductor of the power grid is at zero potential since for the total field distribution of the circuit somewhere in a transformer we have

$$\nabla \times \vec{E} \neq 0 .$$

Or consider an electronic device: When it runs on battery, a potential field would exist, but when it is connected to the mains, there would be none.

Origin:

In electrodynamics we like to operate with simple systems like point charges, dipoles, solenoids etc.

Disposal:

Who believes that a rule is important may proceed as in [1]:

“Outside of the conductor a multivalued magnetic potential exists; for the calculation of the field this ambiguity does not play any role.”…“Inside of the conductor there is no magnetic potential.”

Who estimates that this approach is too fussy, will not give so much importance to the subject. In particular he will not formulate a statement like: “For an induced electric field a potential cannot be defined.” Whether a description with a potential is possible will be decided with regard to the space domain that is relevant for the particular problem.

[1] Bergmann-Schaefer, Lehrbuch der Experimentalphysik, Band II, Elektrizität und Magnetismus, Walter de Gruyter, Berlin 1971, p. 176

Friedrich Herrmann

3.21 Induced emf

Subject:

From three physics textbooks for the secondary school:

“A change of the current intensity induces an electric field within the coil that drives the electric charges…”

“A change of the current intensity I in a coil or the change of the magnetic flux density B which is proportional to it creates within the coil that generates the field an induced electric field strength E_{ind} and thereby an emf V_{ind}, that acts against the change.”

“When the electric current intensity in the large coil is changed, the magnetic field strength also changes and in the small coil an emf is induced.”

From a University physics textbook:

“Kirchhoff’s laws are also valid for alternating currents:

1. Junction rule: At any junction the sum of currents flowing into that junction is equal to the sum of currents flowing out of it.

2. Mesh rule: Any mesh, i.e. any closed loop of a circuit, has the total voltage zero. In other words: The voltage between two points of the circuit is the same, whatever the branch of the circuit on which it is measured…”

Deficiencies:

The values of most physical quantities refer to one of the following geometric entities: a point (example: temperature), an oriented line (example: voltage), an oriented surface area (example: force) or a region of space (example: mass). (A quantity that refers to a more complicated entity, as for instance electric resistance or capacitance actually represents an abbreviation for a particular characteristic or relationship: the U-I characteristic, or the Q-U characteristic.) Whenever the value of a physical quantity is given it must be clear to which point, line, surface or space region it belongs: The temperature at point P is 20 °C, the mass of body B (within the space region occupied by B) is 500 g, the force at cross section S of the string is 40 N.

In our citations this rule is not respected neither in the case of the voltage nor in the case of the electric field strength. Both quantities are vaguely attributed to a coil, but this is not the correct assignment.

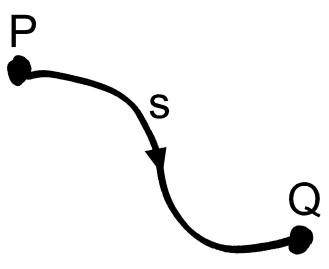


Fig. 1. A voltage or emf refers to an oriented path.

Let us first consider the specification of an emf in general. It is defined by the integral of the electric field strength over a given oriented path,

$$U = \int_P^Q \vec{E} d\vec{r} \ ,$$

for instance path s from point P to point Q in figure 1. The path may also be closed. If the considered field is a conservative field, the value of the emf only depends on the position of points P and Q. If ϕ_P and ϕ_Q are the electric potentials at points P and Q the emf is $U_{PQ} = \phi_Q - \phi_P$. In this case we can say that the emf belongs to the ordered pair of points (P; Q). The value of the emf U_{QP} belonging to the pair (Q; P) has the opposite sign, i.e.

$U_{QP} = - U_{PQ}$.

Often one is only interested in the absolute value of an emf. It has become common practice to speak in this case of “the voltage between points P and Q” without mentioning an order, in the same way as one speaks of a “distance between two points” and thereby means a positive value. This is acceptable as long as one is aware of the wrong conclusions that might be drawn.

Let us come back to the statements of our citations. We consider an R-L circuit with a coil of only one winding, Fig. 2. The emf around a closed path is

$$\oint \vec{E} d\vec{r} = -\iint \dot{\vec{B}} d\vec{A} \ .$$

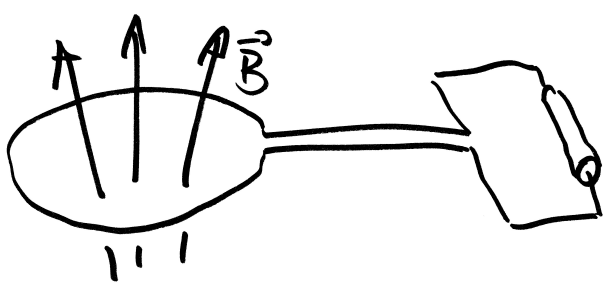


Fig. 2. On the closed path of the R-L circuit only the resistor contributes to the integral over the electric field strength.

As the integration path we chose a line that follows the electric circuit; it runs inside of the electric wire and inside of the resistor. The orientation of the surface element $d\vec{A}$ and the path element $d\vec{r}$ are related by the right-hand rule: If the thumb of our right hand points in the direction of $d\vec{A}$, the bended fingers point into the direction of $d\vec{r}$, Fig. 3. Thus, it does not matter how we orient $d\vec{A}$ as long as we orient $d\vec{r}$ correspondingly. The sign of the induced emf refers to the integration path that was defined in this way.

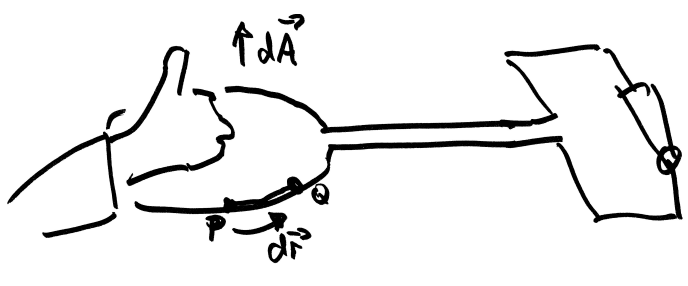


Fig. 3. The orientation of the surface element and the path element in Maxwell’s second equation are related by the right-hand rule.

We now ask how the various parts of the path contribute to total emf. In other words: Which values does the integral

$$U = \int_P^Q \vec{E} d\vec{r}$$

have for various choices of points P and Q? We keep in mind that the integration always runs in the direction of $d\vec{r}$, i. e. Q is ahead of P.

For each section P-Q of the conductor Ohm’s law holds:

$U = R \cdot I$.

We suppose that the resistance of the wires of our “coil” and the feed cables can be neglected compared with the resistance R_0 of the resistor. This means that the voltage on all sections PQ of the path outside of the resistor is zero. Thus, only the resistor contributes to the integral over the total closed path. Only within the resistor is the electric field strength different from zero. Only within the resistor an emf is needed to sustain the electric current.

The arguments remain essentially the same when instead of the single loop we consider a coil with more windings. Then the total magnetic flux through the circuit is N times the flux through one turn and the integration surface has a somewhat intricate shape. But the result is the same: Only the section within the resistor contributes to the total emf.

Consider now our citations 1, 2 and 3: There it is said that the emf is induced *within* the coil. But what could be meant with that? To specify the emf unambiguously the path to which the emf refers should have been indicated. When saying “within the coil” one suggests that the integration path runs somewhere inside the coil. However, the contribution to the integral on any path inside the coil is zero. Thus, any emf within the coil is zero.

This awkwardness is not only found in books for the secondary school. Our forth citation is from a university text book. Here, the mesh rule is used to establish the differential equation for an oscillating circuit. It is even insisted that the emf must be measured along a path. It is not true, however, that the same value of the voltage results on both branches of the circuit. The reason is that the mesh rule is no longer valid as soon as the magnetic flux through the circuit changes with time (just as the junction rule does not hold anymore as soon as electric charge accumulates at the junction).

Origin:

When teaching electricity, the voltage is usually introduced in the context of conservative fields. In this particular case a voltage can indeed be attributed to a pair of points. The path of the line that connects both points does not matter. When trying to describe electromagnetic induction with this habit, one necessarily runs into difficulty.

An additional difficulty arises from the conviction that voltage can be defined operationally by referring to the voltmeter: “A voltage is what is measured with the voltmeter.” The voltmeter seems to be correctly employed as soon as its terminals are connected in any way with two points P and Q of the circuit. However, in the case of the circuit of figure 2, any voltage can be measured between two given points P and Q, depending on how the wire leads are installed. (Notice, that the connecting wires themselves can be wound up to form a coil.)

Disposal:

When dealing with a conservative field the emf is equal to a potential difference. In this case a voltage is defined when an ordered pair of points is specified. However, indicating the order of the points makes the indication a bit clumsy, and actually it is rarely done. One avoids this clumsiness, when describing the circuit from the beginning with potential values instead of potential differences. However, when electromagnetic induction is involved such simplifications are no longer allowed. According to the kind of students we propose one of two procedures:

University: Whenever giving the value of a voltage or emf, specify the oriented line to which it refers. This line can have a beginning and an end, or it can be closed. If it is closed and if it coincides with the conductors of an electric circuit, then it is sufficient to specify the circuit.

High school: Here we usually deal with induced emf’s in coils (not in a single closed wire loop). Thus, do not make statements about the coil but only about the remainder of the circuit or sections of the remainder of the circuit. With them we can deal as we are accustomed from a DC circuit: We can attribute an electric potential to each of its points.

Friedrich Herrmann

3.22 Eddy currents

Subject:

In the scientific and technical literature as well as in physics textbooks for the school, the concept of *eddy currents* is introduced. The following definitions which are taken from the literature, try to explain what is meant by an eddy current:

1. “An induced emf not only appears in a conductor loop and a coil. Electromagnetic induction also takes place when the magnetic field changes within a massive metallic body. Due to the extension of the body circular currents appear, which are called eddy currents.”

2. “Such induced currents within metals do not follow a well-defined pathway as would be the case in a wire or a coil. We call them eddy currents.”

3. “Among the induction phenomena there are emf’s and thus also currents whose path is seemingly disordered. These currents generate magnetic fields, that act against the direction of movement and thus hinder the movement. These currents are called eddy currents.”

4. “If a transformer has a core of massive iron, such currents also appear within this core and heat it up […]. These currents are called eddy currents.”

5. “If the disk moves within the homogeneous *B* field, the field that crosses each piece of metal is changing. At its circumference a circular emf is induced. The corresponding circular currents or *eddy currents*, that flow everywhere within the metal, experience Lorentz forces.”

6. “…They are called eddy currents because the stream-lines of the induced currents are closed like an eddy. Eddy currents create a magnetic field that is opposed to its cause, i.e. the original magnetic field, according to Lenz’s law.“

7. “Eddy current: The alternating current that is induced within an electric conductor by an alternating magnetic field or by the movement in an inhomogeneous magnetic field. The heat produced thereby (Joule effect) can be used to melt metals (induction furnace). In general it appears as an unwanted power loss (eddy current losses), which is diminished in transformers by making the core of a stack of plates that are insulated from each other. Application: in eddy current brakes, for the damping of electric measuring instruments and for the creation of a couple of forces in the AC electricity meter.”

Two more quotations in the context of superconductivity in which the term eddy current is *not* applied:

8. “An external magnetic field induces a circular current, that causes an opposing magnetic field inside the superconductor, that compensates the external field.”

9. “The same is true for external magnetic fields. These induce a circular current, which completely squeezes the magnetic field out. Kamerlingh Onnes started a circular current in a coil and switched the battery off.”

Deficiencies:

It is not easy to understand what is actually the characteristic of an eddy current.

Most definitions emphasize that they are closed or circular currents. This, however, is also true for other currents, even those in a circuit with a battery. It is not the case only if $d\rho/dt$ (the time rate of change of the charge density) is unequal from zero somewhere in the circuit.

Definition 2 stresses that for an eddy current the pathway is not well-defined. Should that mean that the current could take another path without any reason or cause? Any current flows (in an isotropic conductor) in the direction of the electric field vector. The electric field lines define the path of the electric current. This is true for a current in a wire just as much as for the “eddy” current in the core of a transformer or an eddy current brake.

Quotation 3 says that an eddy current is seemingly disordered. What is meant by that? Is it meant that we cannot know the path of the current or is it meant that we simply did not make the effort to calculate the current distribution?

Quotation 1 and 4 emphasize that the body in which the eddy current is flowing is a massive body. What is meant by massive? Simply a great extension? But sometimes wires are rather massive and eddy current brakes tiny.

In some of the definitions reference is made to the effect of an eddy current. In references 4 and 7 heat production is mentioned, and in 3 and 7 the braking property that is due to Lenz’s rule. However, currents that are not called eddy currents also produce heat and they also have the braking effect. Each generator suffers this braking effect.

Quotation 7 shows best, that the definition is the same as that of any induced electric current.

Finally a comment regarding the term circular current in quotations 8 and 9. What is meant is that the stream lines of the current have no ends. However, this is true for any circuit which is not interrupted by a capacitor. We normally do not emphasize that the current in a circuit is circular. This is evident already from the fact that we use the word “circuit”.

We can state that one and the same phenomenon, is given another name, according to the circumstances in which it appears. In the brake we have an eddy current, in a superconductor a circular current and in the pressing iron or the light bulb we have a common current without a name.

It is useful to differentiate by giving proper names to different phenomena if the names grasp an essential characteristic. Otherwise it is counterproductive. Then it is better to use a unique wording in order to point out the similarities.

The term eddy current and the statement that the current distribution is disordered or undefined may cause yet an additional problem. It strongly suggests a similarity with turbulent flows of liquids, i.e. a phenomenon where the terms disordered and undefined are appropriate. They are, however, in this respect fundamentally different from the eddy currents of electricity.

Origin:

In 1824 François Arago discovered that a magnetic needle that could freely rotate was dragged by a rotating copper disc. This observation led Faraday in 1832 to the discovery of electromagnetic induction.

A quickly rotating copper disc that is placed between two magnetic poles is slowed down. Foucault concluded in 1855 that the work that has been spent to set the disc into rotation must reappear as heat. He was able to show it in impressing experiments. In french language eddy currents are called *courants de Foucault*. Both the braking effect and the heat production are since then considered the essential characteristics of an eddy current, although both phenomena also arise in any other induced current.

Thus, the particular circumstances of the discovery have lead to the introduction of a new term or concept, in particular the fact that the dragging effect had been discovered before Faraday gave a more general explanation of the electromagnetic induction.

The situation is rather similar to that of Lenz’s rule, which also has survived its own generalization [1].

Disposal:

Do not use a particular name for currents that are induced in the iron cores of transformers and in eddy current brakes. In order to distinguish the “eddy current brake” from a mechanical brake you may call it induction brake. If there are doubts about if a current is closed, do not call it a circular current, but simply say that the circuit is closed.

[1] *F. Herrmann:* Lenz’s law, article 3.10

Friedrich Herrmann

3.23 Permeability

Subject:

For the magnetic flux density B in a long electromagnet (length l, number of turns N, relative permeability of the core material μ_r) the following formula can be found in text books:

$$B = \mu_0 \cdot \mu_r \cdot \frac{N \cdot I}{l} \qquad (1)$$

It is supposed that the electric current intensity is not too high, because otherwise the core material might approach saturation.

In some text books an equation is found that is equivalent to (1):

$$B = \mu_r \cdot B_0 \ . \qquad (2)$$

Here, B is the flux density within the core material and B_0 is that inside the empty coil, i.e. the coil from which the core has been removed.

Deficiencies:

Equation (1) and (2) are not correct. They are valid only if the whole space which is occupied by the field is filled with the material of permeability μ_r.

"In those cases where a homogeneous and isotropic magnetizable substance occupies the whole space of the magnetic field, or a part of it in such a way that the lines of induction of the magnetizing field do not traverse the surface of the magnetized material, inside of the material the relation

$B = \mu_r \cdot B_0$

holds, where μ_r is the relative magnetic permeability of the magnetizable substance, which…". [1]

Thus, equations (1) and (2) are valid for instance for a toroidal coil with a closed core.

In order to understand why the equations are not valid for a normal straight electromagnet, let us begin by deriving the correct expressions that correspond to equations (1) and (2) for the case of a coil with a toroidal core that is provided with a slit, Fig. 1. Thereafter we consider the case of a stretched coil.

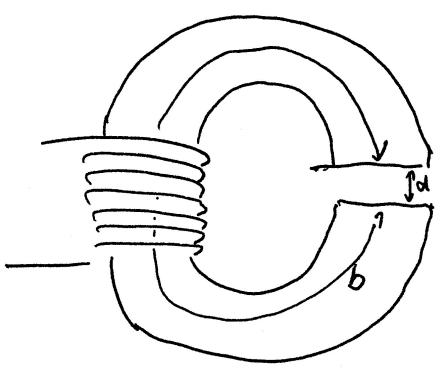


Fig. 1. The flux density inside the iron core and in the slit is proportional to μ_r only if the slit width d is sufficiently small.

We suppose the slit width to be so small that the field within the slit can be considered homogenous.

Since we are far from saturation and since the material is isotropic, we have everywhere

$$B = \mu_0 \cdot \mu_r \cdot H \ . \qquad (3)$$

Now admit that the coil has N turns and the electric current is I. We then have:

$$\oint \vec{H} d\vec{r} = N \cdot I$$

where the integration is over a path that follows the torus.

We now suppose that the radius of our ring (the great radius of the torus) is great compared with the radius of a cross section of the ring (the small radius of the torus). We now can easily evaluate the Integral:

$$b \cdot H_m + d \cdot H_s = N \cdot I \qquad (4)$$

The index m refers to the material of the core, s refers to the slit. b is that part of the integration path that runs inside the core material, whereas d refers to the path section inside the slit.

Since the B field is divergence-free we have

$B_m = B_s = B$,

i.e. B is the same in the material and in the slit.

We then get by using equation (2):

$\mu_r H_m = H_s$.

Inserting in equation (4) we get

$$H_m = \frac{N \cdot I}{b + \mu_r d}$$

and by using equation (3)

$$B = \frac{\mu_0 \cdot \mu_r \cdot N \cdot I}{b + \mu_r d} \ . \qquad (5)$$

We see that the dependence of B on μ_r is not that of equation (1).

For an empty coil we get

$$B_0 = \mu_0 \cdot \frac{N \cdot I}{l} \ ,$$

where l is the total integration path length. With this equation (5) becomes

$$B = \frac{\mu_r \cdot l}{b + \mu_r d} \ B_0 \, .$$

It is seen that the relation between B and B_0 is not that claimed by equation (2).

Let us now evaluate equation (5) for two special situations:

1. if the solenoid does not have any slit, i.e. if $d = 0$, or also approximately if $b \gg \mu_r d$ we get

$$B = \mu_0 \cdot \mu_r \cdot \frac{N \cdot I}{b} \ . \qquad (6)$$

In this approximation the flux density is independent of the (small) slit width and proportional to the relative permeability of the core material. Since the path b inside the material is (almost) equal to the total path l, the equation turns out to be identical with equation (1). Thus, we see that equation (1) is valid only when the core does not have a slit or if the slit is small compared with b/μ_r .

2. If $b \ll \mu_r d$, equation (5) becomes approximately

$$B = \mu_0 \cdot \frac{N \cdot I}{d} \ . \qquad (7)$$

Now B is independent of μ_r, but inversely proportional to the slit width. By comparing with the empty coil we get

$$B = \frac{l}{d} B_0 \ . \qquad (8)$$

Also in this case B is independent of μ_r.

Now, which of these two approximations correspond to the situation that we meet at school?

Suppose we have a ring magnet of a total length of 50 cm, that would result if constructed from typical school material. Further suppose that we provide the magnet with a slit not smaller than 0.2 cm (otherwise it would not be possible to introduce the Hall probe). If we admit that μ_r = 1000 we find that b in the denominator at the right side of equation (5) is only one forth of $\mu_r d$. We are thus in the scope of validity of equations (7) and (8), and not that of equation (6).

Actually at school the measurement of μ_r is not made with a ring magnet but with a straight solenoid. In this case the distance between the two poles of the iron core is even much greater. So one is definitely in the range of validity of equation (7). Indeed, these experiments give a value of μ_r that is too small by more than a factor of 10.

Equations (1) and (2) suggest that the magnetic flux density inside the core of an electromagnet increases in proportion to the specific permeability of the material. This would mean that the flux density increases by the same factor outside at the surface of the core. Then an electromagnet with a core with μ_r = 100 000 would have a flux density that is a hundred times that of a core with μ_r = 1000. However, equation (7), and also common sense tells us that this cannot be true. An electromagnet with μ_r = 500 can hardly be improved by choosing another core material. (This does not mean that in some situations a material with a very great μ_r is not indispensable.)

Origin:

We have found the incorrect equations in all of the five high school text books that we have consulted, but not in any university text book or encyclopedia. This gives a hint on how the error originated. School physics has to get along with as few physical quantities as possible. So one tries to introduce the specific permeability without using the magnetic field strength H. The wrong conclusion might have been, that this can be done in a way that is analogous to introducing the dielectric constant ε_r . This is done by inserting a dielectric into the space between the plates of a capacitor and measuring the decrease of the potential difference between the plates. The fields strengths with and without the dielectric material are related by

$$E_0 = \varepsilon_r \cdot E \, .$$

This relation is applicable, in contrast to its magnetic counterpart, since in the case of the capacitor the whole space that is occupied by the field is filled with the dielectric. And in addition for common dielectric materials ε_r is much smaller than μ_r for typical iron core materials.

Disposal:

The description of magnetostatic phenomena gets clearer when employing H instead of B. Doing so we can formulate a simple rule:

A softmagnetic material displaces the magnetic field (measured by H) from its inside in the same way as an electric conductor displaces the electric field.

For most applications it doesn't make any difference whether the field is, according to the value of μ_r, displaced to 99,9 % or 99,99999 %.

[1] *B. M. Jaworski* and A. A. *Detlaf*: *Physik griffbereit*, Vieweg, Brauschweig 1972, p. 410

Friedrich Herrmann

3.24 Ignition spark and electromagnetic radiation

Subject:

Everybody knows that electric sparks cause radio and TV interference. During a thunderstorm, when a switch is operated or when an electric motor is running, one can hear a cracking noise when receiving an amplitude modulated signal. The ignition sparks of a car engine would also cause interference if the car had not interference suppression circuitry implanted. In the original version of Hertz's experiment for showing the existence of electromagnetic waves, sparks play an important role.

It is a wide-spread opinion that the radiation that causes the interference originates from the spark gap:

"The spark generates high frequency interference pulses, which have to be suppressed. In order to do so various measures are possible…"

"…together with the spark electric discharges are arising in the form of electric oscillations; the spark jumps from one sphere to the other; thus, the spark gap between the spheres acts as an emitter."

"An oscillating electromagnetic perturbation (i.g. a spark discharge) generates electromagnetic waves, which propagate with the velocity of the light."

One can find illustrations of Hertz's experiment where, electromagnetic waves are drawn that emanate from the spark gap between the two halves of the oscillator.

Deficiencies:

It is not the spark gap that emits the electromagnetic radiation but the electric conductor of which the spark gap is only a very small part. In the case of the Hertz oscillator the whole antenna is emitting. The role of the spark gap is that of a switch which connects the two parts of the antenna as soon as the voltage has attained a certain value.

This voltage has to be very high, in order to get a high electric field strength, and in order to get a high magnetic field strength after closing the switch. When charging, the halves of the antenna must be disconnected. Instead of connecting them by means of a normal switch one uses the much simpler method of the spark. As soon as the discharge is initiated there is a conducting connection. Even though the voltage passes through zero as the oscillation takes place, the spark does not cease, since the ionization of the air survives.

The same is true for the sparks of a light switch or of the brushes of an electric motor: The emission of the electromagnetic wave does not occur only at the spark, but at the whole of the conductor in which the current is fluctuating when the circuit is opened. Thus, the spark is a necessary condition for the occurrence of the emission of the waves, but the source of the wave is the entire conductor in which the rapid change of the current takes place.

Origin:

Everybody knows: When there is a spark, there is also the cracking noise. The spark is eye-catching, there is light and sound coming from it. It seems plausible that the spark is also the source of the electromagnetic wave that causes the interference.

The misconception survives although it is in contradiction with what the students learn about the dipole antenna: It is the whole antenna which is responsible for the emission.

Disposal:

Explain clearly that the role of the spark is only that of an automatic switch. The spark gap establishes a conducting connection between two metallic conductors.

Friedrich Herrmann

3.25 Mechanical stress within the electric and within the magnetic field

Subject:

Electric and magnetic fields are under mechanical stress. This stress is strong and it is easy to feel it. It can be calculated with a simple formula. Nevertheless, in the text books for the High school and the University it is hardly ever mentioned. Sometimes it is dismissed as fictitious stress.

Deficiencies:

When treating electric and magnetic forces without mentioning the stress in the fields, they necessarily appear as actions at a distance. When saying that the plates of a capacitor attract each other without mentioning the tensional stress within the field between the plates, one only can conclude that the force of one plate on the other is acting over a distance. At least since the birth of Maxwell's electrodynamics no physicist believes anymore in such actions. Newton already considered such forces a senseless idea, and Maxwell pointed out at various occasions that for him an „actio in distans" is absurd.

Origin:

The problem is that the force concept which we use still today suggests actions at a distance. Newton had introduced the "force language" only reluctantly. For him there was no other way, since the concept of field did not yet exist. When finally Faraday and Maxwell introduced the field concept it was already too late [1]. Newton's stopgap solution (body A exerts a force on body B without any intervention of an intermediate medium) had already become the doctrine. However, there was yet another misfortune. With the banishment of the ether from physics fields became no more than a chimera – hardly more than a mathematical tool for calculating forces. An entity that is under mechanical stress was far too concrete to fit in the field idea of the etherless interregnum. When space began to fill up again after the appearance of the General Theory of Relativity and Quantum Electrodynamics, it was far too late. Mechanical stress within the field remained "fictitious", or worse: it was completely ignored.

Disposal:

Take fields as objects seriously. It has all the standard properties that we know from material systems – only to another degree. The mechanical properties of a field are easy to calculate from the field strengths. For the electric field the tensional stress in the direction of the field lines is

$$\sigma_{\parallel} = -\frac{\varepsilon_0}{2}|\boldsymbol{E}|^2$$

The compressional stress perpendicularly to the field lines is

$$\sigma_{\perp} = +\frac{\varepsilon_0}{2}|\boldsymbol{E}|^2 .$$

The absolute values are equal to the energy density of the field. The expressions for the magnetic field are analogue: instead of ε_0 there is μ_0, and **H** stands for **E**.

In order to help to get a concrete idea of the field, when teaching I mention the fact that one liter of the magnetic field of a neutron star has a mass of 1 kg and that the weight of this liter on the neutron star is $2 \cdot 10^{11}$ the weight of 1 kg on the Earth.

[1] *J. C. Maxwell*: A treatise on Electricity and Magnetism, Volume One, Dover publications, New York, 1954, Article 105, p. 157: "If the action of E_2 on E_1 is effected, not by direct action at a distance, but by means of a distribution of stress in a medium extending continuously from E_2 to E_1, it is manifest that if we know the stress at every point of any closed surface s which completely separates E_1 from E_2, we shall be able to determine completely the mechanical action of E_2 on E_1."

Friedrich Herrmann

3.26 Closed *B* field lines

Subject:

It is often said that magnetic field lines are closed:

"The […] difference is that electric field lines always begin on positive charges and end on negative charges, whereas for magnetic field lines there are no points in space, where they begin or end, since magnetic monopoles do not exist. Instead magnetic field lines form closed loops."

"The magnetic field of a current has always closed force lines, in contrast to the electrostatic field lines, which start on positive charges (sources) and end on negative charges (sinks)."

Deficiencies:

1. Normally, when referring to magnetic field lines, the field lines of the magnetic flux density **B** are meant. The fact that field lines can have a beginning and an end is not a peculiarity of the electric field. Just as the **E** field lines of an electrostatic arrangement have a beginning and an end, the **H** field lines of a magnetostatic arrangement have a beginning and an end: They begin on a north pole of a magnetized body and end on a south pole.

2. The fact that the **B** field is divergence-free does not allow for the conclusion that the **B** field lines are closed, and indeed, in general they are not. The equation

div **B** = 0

only tells us, that the field lines do not have a beginning and or an end.

What do we mean in the first place when we say, field lines are closed? Probably anybody who hears the statement will imagine something like the following: We have an electric current that is flowing on a well-defined path, typically within a wire. An arbitrarily chosen field line runs around this current. When beginning at one point of the line and following the line we come back to our starting point after one single turn around the current.

However, there is no physical reason why the line should close after one turn. And practically there is only a vanishing small chance that this would happen. If exceptionally the line does so, the reason is not so much physical but rather geometrical. It is indeed so in the case of a straight wire of infinite length, or in the case of an electric circuit that is completely confined to a plane. The smallest deviation from this restriction makes that the line, after executing one turn misses its point of departure. One might believe that the field of a cylindrical or toroidal solenoid has closed field lines, but they don't [1]. The unavoidable helicity of the coils is the reason why a field line does not meet its starting point after one turn.

From Fig. 1, which shows a straight current-carrying wire and a circular current, it can be seen that field lines are in general not closed. We consider the field vectors in the plane of the ring. At the inner side of the ring the superposition of the fields of the circular current and of the straight conductor result in a left-hand helix; outside of the ring it is a right-hand helix. Obviously, the field lines cannot be closed.

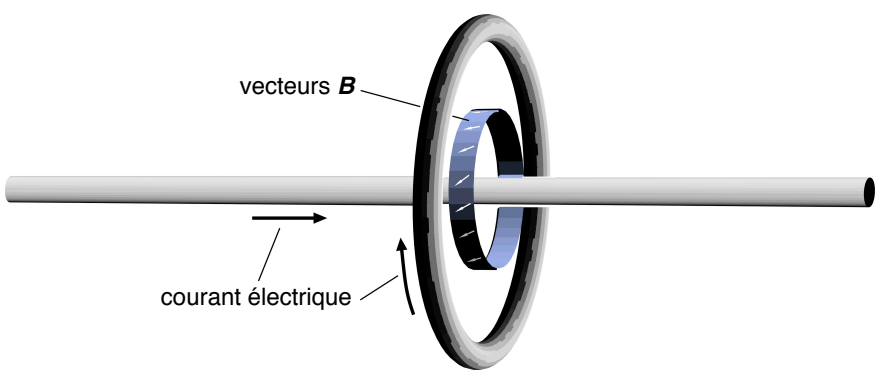


Fig. 1. Straight current and circular current. The field lines in the plane of the ring form a left-hand helix inside and a right-hand helix outside of the ring.

Fig. 2 shows another example of field lines that are not closed. A cylindrical homogeneously magnetized flexible permanent magnet is twisted around the axis of the cylinder, and then bended and closed to form a torus. The lines of magnetization, and thus the **B** lines now spiral round the torus axis (which previously was the axis of the cylinder) and never close.

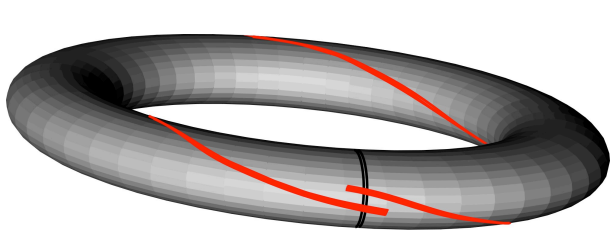


Fig. 2. A flexible permanent magnet which initially was cylindrical has been twisted and bent.

Magnetic fields in nature, for instance those of the Earth or cosmical magnetic fields are so intricate that nobody would suspect that field lines might ever be closed.

Another technical example where a field line after running around a current misses its point of departure by far is the magnetic field of a fusion reactor.

It was shown in a beautiful article in the American Journal of Physics in 1951 by Joseph Slepian [2] that **B** field lines are in general not closed. The article does not contain a single equation or figure. In the following decades several other articles were written about the subject, see [1, 3] and the literature which is cited there.

Origin:

1. At school and in the University lecture we usually only treat simple magnetic fields: the field of the bar magnet and of the horse shoe magnet, the field of electric currents in a straight conductor or in a conducting ring, and the field in a solenoid. The **B** field lines of an ideal bar magnet, i.e. a bar magnet with a perfectly homogeneous magnetization, or the field lines of a perfect horseshoe magnet are indeed closed; the same is true for the field of a perfectly straight wire or a perfect circular conductor. In the case of a solenoid it is true only approximately. The fact that mainly these sources of magnetic fields are considered may explain why it seems plausible that the field lines are always closed.

2. Field lines are a graphical tool for the representation of field strength distributions. However, students often perceive them as something to which a physical reality can be attributed. If one imagines the field lines as physical objects, there is an argument in favor of the idea that field lines are closed, even if it is not after one single turn around a current carrying conductor. Instead of a field line consider a thread. Someone has made of the thread a mazy clew and assures us that the thread has no beginning and no end. In this case the conclusion that the thread forms a closed loop or several closed loops is correct. Why does this argument not hold for magnetic field lines? Field lines are no physical objects but mathematical objects, i.e. lines. All one could say in the best case is that, when following a field line, one may come as near as one wants to the point of departure.

Disposal:

Avoid saying that field lines are closed. It is enough to say that they have no starting point and no end. However thereby one would only try to eliminate a symptom; the roots of the evil are deeper. The proper cause of the error is the misconception of the field lines as physical objects.

Therefore, it is more important that when introducing the field concept one does not begin with representing the field by field lines. The picture to show first could rather be a representation of the energy distribution of the field by means of a gray shading, Fig. 3a.

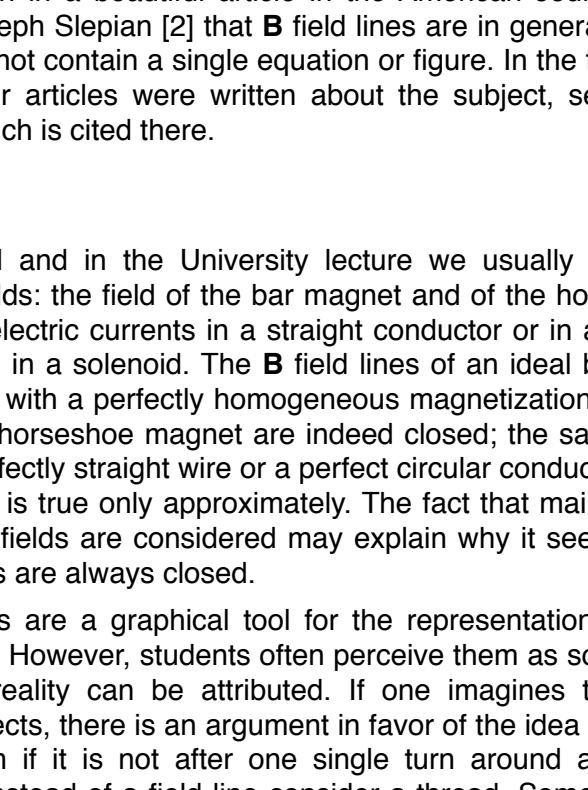


Fig. 3. (a) Representation of the field of a solenoid by means of gray shading; (b) Representation of the field of a solenoid by means of gray shading, field lines and field surfaces

Only thereafter one shows that the field has in every point a preferential direction, i.e. it is not isotropic. To graphically represent this fact one begins by drawing vector arrows. Next one comes to a representation that is more convenient for practical purposes: one draws the field lines, but also the field surfaces, fig. 3b. By field surfaces we mean the orthogonal surfaces to the field lines. For conservative fields they are identical with the equipotential surfaces. For Maxwell it was a matter of course to represent in all of his figures field lines and field surfaces [4]. Both elements have a simple intuitive meaning: In the direction of the field lines the field is under tensional stress, in all the orthogonal directions, i.e. the directions within the field surfaces there is compressional stress. When knowing that, one will no longer interpret the field lines as filaments that run through the field, but as a means to represent graphically the mechanical stress within the field.

[1] *M. Lieberherr*: The magnetic field lines of a helical coil are not simple loops, Am. J. Phys. 78 (2010), S. 1117-1119

[2] *J. Slepian*: Lines of Force in Electric and Magnetic Fields, Am. J. Phys. 19 (1951), S. 87-90

[3] *M. Schirber*: Magnetic Fields in Chaos, Phys. Rev. Focus, http://focus.aps.org/story/v24/st24

[4] *J. C. Maxwell*: Lehrbuch der Electricität und des Magnetismus, Verlag von Julius Springer, Berlin 1883, Tafeln XII bis XXI

Friedrich Herrmann and Ralph von Baltz

3.27 Magnetic monopole and magnetic charge

Subject:

There are no particles that carry magnetic charge. It is said, that no magnetic monopoles exist. It follows, so it is argued, that a physical quantity „magnetic charge“ or „magnetic pole strength“ does not exist.

Deficiencies:

Let us first clarify two concepts:

Magnetic charge density ρ_m:

It is defined by

$$\mu_0 \operatorname{div} H = \rho_m \qquad (1)$$

Thus, the magnetic charge density describes the sources of the H field. Since we have

$$\mu_0 \operatorname{div} H = -\operatorname{div} M \qquad (2)$$

it also signifies the sinks of the magnetization field. The volume integral of the magnetic charge density is called *magnetic pole strength*, *magnetic charge* or *amount of magnetism*.

Magnetic monopole:

The word is not used consistently.

When it is said that no magnetic monopoles exist, one refers to particles, i.e. objects that carry magnetic charge (or whose pole strength is different from zero). Such „monopoles“ have not been found so far.

But the word is also used as a name for the source of a magnetic „Coulomb field“, i.e. a magnetic field, whose field strength H decreases as $1/r^2$. Such fields can be realized in any desired approximation. It is the field in the vicinity of a pole of bar magnet that is long and thin.

Because of this ambiguity, in the following we shall not call a magnetically charged particle a „monopole“ but a „monopole particle“.

To show the non-existence of the physical magnitude „magnetic charge“ it is usually argued that no monopole particles have been found. However, in order to explain what is meant by such a particle it is necessary to first introduce the physical quantity magnetic charge, for instance by means of equation (1).

It is not possible to deduce the existence or non-existence of a physical quantity from an observation of nature. Physical quantities are human constructs or creations [1]. A physical quantity is introduced when it is advantageous; when it can serve for the description of natural phenomena. Actually it is advantageous to introduce a quantity „magnetic charge“. It is needed among other things:

- to describe the fact, that no magnetic monopole particles exist;
- to describe the fact, that the poles of a magnet carry equal and opposite charges at their poles;
- to formulate Coulomb’s law for magnetic poles [2].

Certainly one could do without the introduction of magnetic charge. But then, instead of saying that no particles exist that carry magnetic charge, we had to formulate: „There are no particles for which the volume integral of the divergence of the magnetic field strength over a region of space that contains the particle is different from zero.“ By the way, one could get rid of the electric charge in the same way. Obviously nobody would do so.

Origin:

Magnetic charge is a time-honored physical quantity, which has been used with various names. At Coulomb’s time it was imagined that magnetism is caused by two magnetic fluids (in analogy with the electric phenomena, which were explained by two electric fluids).

For both, the electric and the magnetic fluids, Coulomb discovered the inverse square law of the force [2].

Maxwell calls this quantity the „strength of a pole“ [3]:

> *The repulsion between two like poles is in the straight line joining them, and is numerically equal to the product of the strengths of the poles divided by the square of the distance between them.*

On the next page he introduces the term „quantity of magnetism“, and he states:

> *The quantity of magnetism at one pole of a magnet is always equal and opposite to that at the other, or more generally thus:*
>
> *In every Magnet the total quantity of Magnetism (reckoned algebraically) is zero.*

The term quantity of magnetism is also later used by Max Born [4].

Although it is easier to verify experimentally Coulomb’s law for magnetic than that for electric charge, the quantity has today almost completely disappeared from the text books. This happened together with the marginalization of the magnetic field strength. If the field strength is not used to describe a magnetic field the equation

$$\mu_0 \operatorname{div} H = \rho_m$$

no longer can serve to get a feeling for the magnetic charge.

Disposal:

At school: Introduce magnetic charge right from the beginning of magnetostatics as an independent extensive physical quantity, in the same way as one introduces electric charge in electrostatics. The total magnetic charge of a magnet is zero.

At university: First treat the relation

$$\mu_0 \operatorname{div} H = -\operatorname{div} M$$

Thereafter introduce magnetic charge as:

$$\rho_m = \mu_0 \operatorname{div} H\,.$$

Friedrich Herrmann

[1] G. Falk und W. Ruppel: Mechanik, Relativität, Gravitation, Springer-Verlag, Berlin 1973, p. 2

[2] C. A. Coulomb: Second Mémoire sur l’Électricitá et le Magnétisme, Où l’on détermine, suivant quelles loix le Fluide magnétique, ainsi que le Fluide électrique, agissent, soit par répulsion, soit par attraction. Mémoires de l’Academie Royale des Sciences, 1785, p. 593

[3] J. C. Maxwell, A treatise on electricity and magnetism, Volume two, Dover Publications, Inc, New York, 1954, p. 3-4

[4] M. Born, Die Relativitätstheorie Einsteins, Heidelberger Taschenbücher, Springer-Verlag, Berlin 1969, p. 133

3.28 Equivalent resistance

Subject:

If several resistors with the resistances R_1, R_2, R_3, … are connected in series the total or equivalent resistance of the system is

$$R = R_1 + R_2 + R_3 + \ldots$$

If they are connected in parallel the equivalent resistance is

$$\frac{1}{R} = \frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3} + \ldots$$

Deficiencies:

These rules are part of the physics syllabus, since electricity is taught at school, i.e. since about 150 years.

Of course, nothing is incorrect. But one might ask several questions: Why do these rules belong to the compulsory part of the syllabus? Why do they have the status of standard rules? Why do we dedicate a whole chapter to them? One might answer: Because they are important.

But if one considers them as sufficiently important to be treated extensively in the physics class, why not also a good number of similar or analogues rules: about connecting capacitors and solenoids in series or in parallel, or Hookean springs and mechanical dampers (dash pots), or heat resistors and fluid resistors? The corresponding mathematical structure of the rules is the same as that for the electrical resistors. Is heat resistance less important than electric resistance? Are electric capacitors less important than electric resistors?

When taking into account that the above-mentioned rules are consequences of the the junction rule and the mesh rule, another observation can be made.

The loop rule becomes trivial, when the voltage is introduced as a difference of two electric potential values. Then, the mesh rule is as trivial as for instance the following statement: If taking the elevator one goes up first two floors and then three more, one has gone up five floors in total. One can say: the mesh rule is valid because we have to do with a conservative field. (But it is better not to express it in such an intimidating way.)

The junction rule is a simple consequence of the conservation of the flowing quantity. It is valid for any flow of a conserved quantity: energy currents, electric currents, mass currents and momentum currents. It would be a pity when treated only in the context of electric charge.

Origin:

The rules had been formulated (in a form that is slightly different from ours) in 1845 by Kirchhoff, i.e. at the beginning of electricity. At that time everything was new, and they appeared not trivial at all. The reason why they survived until the day of today may be due to the fact that they had gotten their own proper name, namely Kirchhoff's laws (in contrast to the above-mentioned analogues rules)..”

Disposal:

There is nothing to say against treating the various ways of wiring electric resistors as one of many other problems of electricity. However, one would not give the results the status of rules or laws. And one will treat similar questions related to other devices, like capacitors and inductors, and with other currents, like energy, momentum, heat and water currents.

Finally, one of my favorite rules for writing a syllabus: Always when someone proposes to introduce a new subject into the syllabus (or into the curriculum or into a text book) look first for competing subjects, i.e. subjects that are similar due any kind of analogy. Only if you find an argument, to treat the subject that was first proposed, and not its competitors, the subject is accepted. This method has proved to be useful in many occasions.

Friedrich Herrmann

3.29 Symmetries in electromagnetism

Subject:

Electromagnetism is rich in internal structures, symmetries or analogies. They manifest in the phenomena and become visible in the theoretical description. Some examples:

- Coulomb has discovered the law that carries today his name both for electric charges and magnetic poles.
- On a moving body that is electrically charged two forces are acting that often are presented as being analogs: one is proportional to the electric field strength *E*, the other, the Lorentz force, is proportional to the magnetic flux density *B*.
- In electrical engineering the capacitor and the coil, and thus capacitance *C* and inductance *L* play analog roles. This is seen for instance in the oscillating circuit.

Deficiencies:

The characteristic of the structures that we are considering here is that physical quantities, as well as mathematical relations between them can be mapped. When replacing the quantities in one equation according to certain given rules that characterize the analogy one gets a new relation that is also correct.

A problem that sometimes is not seen is that within a given physical domain there may exist several competing analogies, that are based on different mappings. An example is electromagnetism. The above mentioned examples have to do with such competing analogies. Who is not aware of the fact that various analogies exist, risks to run into difficulties. Which is the magnetic analog to the electric field strength *E*? Is it *B* or *H*? Sometimes it seems to be the one, sometimes the other. It even seems that sometimes ideological points of view are showing up: the „actual" or „true" magnetic field is *B* (or *H*). There are school physics books where without further ado the magnetic flux density is rebaptized magnetic field strength.

The problem resolves itself, if one realizes that we have to do with more than one mapping. Then the question is no longer: Which is the correct correspondence but which is the one that is convenient in view of a given question.

Let us remind the three analogies within electrodynamics by means of three tables. Each of these tables contains the physical quantities that correspond to one another, as well as some of the relations between them.

1. The analogy $\vec{E}$ - $\vec{H}$

It manifests itself in Maxwell's equations, table. 1.

electric vector field quantities	*magnetic vector field quantities*
electric field strength $\vec{E}$	magnetic field strength $\vec{H}$
electric flux density $\vec{D}$	magnetic flux density $\vec{B}$
polarisation $\vec{P}$	magnetization $\vec{M}$
electric charge	*magnetic charge*
total charge Q	total charge Q_m
charge density	*charge density*
of total charge ρ	of total charge ρ_m
of free charge ρ_F	of free charge $\rho_{mF} = 0$
of polarization charge ρ_P	of polarization charge ρ_{mP}
first Maxwell equation	*second Maxwell equation*
$\mathrm{div}\ \vec{D} = \rho_F$	$\mathrm{div}\ \vec{B} = \rho_{mF} = 0$
$\mathrm{div}\ \vec{P} = -\rho_P$	$\mathrm{div}\ \vec{M} = -\rho_{mP}$
$\varepsilon_0\ \mathrm{div}\ \vec{E} = \rho_F + \rho_P$	$\mu_0\ \mathrm{div}\ \vec{H} = \rho_{mF} + \rho_{mP} = \rho_{mP}$
electric current density	*magnetic current density*
conduction current $\vec{j}_L$	conduction current $\vec{j}_{mL} = 0$
displacement current $\vec{j}_V = \dot{\vec{D}}$	displacement current $\vec{j}_{mV} = \dot{\vec{B}}$
third Maxwell equation	forth Maxwell equation
$\mathrm{rot}\ \vec{E} = -\vec{j}_{mL} - \dot{\vec{B}} = -\dot{\vec{B}}$	$\mathrm{rot}\ \vec{H} = \vec{j}_L + \dot{\vec{D}}$
force laws	*force laws*
$\vec{F} = Q \cdot \vec{E}$	$\vec{F} = Q_m \cdot \vec{H}$
$\lvert\vec{F}\rvert = \frac{1}{4\pi\varepsilon_0} \frac{Q_1 \cdot Q_2}{r^2}$	$\lvert\vec{F}\rvert = \frac{1}{4\pi\mu_0} \frac{Q_{m1} \cdot Q_{m2}}{r^2}$
energy current density $\vec{j}_E = \vec{E} \times \vec{H}$	

Table 1. Analogy, in which $\vec{E}$ and $\vec{H}$ correspond to one another. The free electric charge and the electric conductive current do not have a magnetic analogue.

It is particularly helpful when treating electromagnetism. The $\vec{H}$ field has sources and sinks and therefore drawing an $\vec{H}$ fieldline picture is as simple as sketching an $\vec{E}$ fieldline picture in electrostatics. It is well-known that students (and not only they) have difficulties in drawing magnetic field line pictures [1].

2. The analogy $\vec{E}$ - $\vec{B}$

It is based on the representation of electromagnetism with four-vectors. Here, the temporal component of a four vector corresponds to the spacial components. Just as time and position, or energy and momentum, also the electric charge density and the current density, as well as the electric potential Φ and the magnetic vector potential $\vec{A}$ correspond to one another. The spacial derivatives of the potentials, i.e. the gradient of Φ and the curl of $\vec{A}$ result in the vector quantities $\vec{E}$ and $\vec{B}$, respectively, with which the force laws are formulated, table 2.

sources of the electric field	*sources of the magnetic field*
electric charge density ρ	electric current density $\vec{j}$
electric vector field quantity	*magnetic vector field quantity*
electric field strength $\vec{E}$	magnetic flux density $\vec{B}$
potential quantity	*potential quantity*
electric potential ϕ	magnetic vector potential $\vec{A}$
derivative of potential	*derivative of potential*
$\vec{E} = -\,\mathrm{grad}\ \phi$	$\vec{B} = \mathrm{rot}\ \vec{A}$
force laws	*force laws*
$\vec{F} = Q \cdot \vec{E}$	$\vec{F} = I \cdot (\vec{s} \times \vec{B})$ (Lorentz force)
$\lvert\vec{F}\rvert = \frac{1}{4\pi\varepsilon_0} \frac{Q_1 \cdot Q_2}{r^2}$	$\lvert\vec{F}\rvert = \frac{\mu_0}{2\pi} \frac{I_1 \cdot I_2}{r} \cdot \mathrm{length}$ (two parallel currents)

Table 2. Analogy, in which $\vec{E}$ and $\vec{B}$ correspond to one another.

3. The analogy U-I

It is useful in particular in electrical engineering. It is a variant of the $\vec{E}$ - $\vec{H}$ analogy. However, here the correspondence is made crosswise: The electric charge does not correspond to the magnetic charge, but to the magnetic flux, table 3.

physical quantities		
electric charge Q	⇔	magnetic flux $-\Phi$
voltage U	⇔	electric current I
capacitance C	⇔	inductance L
electric resistance R	⇔	electric conductance G
energy E	⇔	energy E
energy flow P	⇔	energy flow P
equations		
$I = \frac{dQ}{dt}$	⇔	$U = -\frac{d\Phi}{dt}$
$P = U \cdot I$	⇔	$P = U \cdot I$
$Q = C \cdot U$	⇔	$\Phi = L \cdot I$
$E = \frac{C}{2} U^2$	⇔	$E = \frac{L}{2} I^2$
components		
capacitor	⇔	coil
voltage-stabilized power supply	⇔	current-stabilized power supply
topological rules		
series connection	⇔	parallel connection
short circuit	⇔	open circuit

Table 3. Analogy that is important in electrical engineering. Not only physical quantities and equations correspond to one another, but also technical components and topological rules.

This analogy is quite different from the two preceding ones. Here, one electric circuit corresponds to another one, and the translation table has to be used from left to right and from right to left. One sometimes calls this kind of relationship a *dualism*. So one has to replace for instance a voltage by an electric current, and a current by a voltage, or a capacitor by a coil and a coil by a capacitor, or a parallel connection by a series connection and a series connection by a parallel connection.

Actually, it would be logical to relate the electric flux to the magnetic charge. The corresponding technical devices would be a „magnetic capacitor" and a „coil", in whose „wire" a magnetic current is flowing. Since no free magnetic charge and no magnetic conduction currents exist (but only bound charge and displacement currents) both elements are not interesting.

Origin:

The $\vec{E}$ - $\vec{H}$ analogy was the most obvious since it is suggested directly by the classical formulation of Maxwell's theory. With the Theory of Relativity and the description with four-vectors the $\vec{E}$ - $\vec{B}$ analogy came into fashion. The third analogy owes its popularity to the fact that it is very useful in electrical engineering, and in addition, that it is the basis of a far-reaching analogy between electricity and mechanics (where the capacitor corresponds to the elastic spring, the coil to inertial body and the electric resistor the dashpot [2]).

The dispute about which of the two quantities $\vec{H}$ or $\vec{B}$ is more convenient, or which of them represents the „correct" or „true" magnetic field may have two causes:

1. One may know only one of the analogies; the other is unknown and appears suspect.

2. One identifies the physical system „field" with the physical quantity „field strength", i.e. one ignores the fact that physical quantities are human creations or constructions.

It seems that even the great Sommerfeld believed that when choosing one or the other representation the question was which is the correct one, and not simply which one is more appropriate to solve a given problem [3]:

> „The Faraday-Maxwell induction law shows, that the electric field strength *E*, as well as the magnetic induction *B* are quantities that describe an intensity; so *B*, not *H* merit the name *magnetic field strength*. […] It follows unambiguously from the Theory of Relativity that *B* belongs to *E* and *H* to *D*, since the quantities *cB* and *–iE* on the one hand, and *H* and *–icD* on the other are coupled to form a six vector (antisymmetric tensor)." *Disposal:*

1. Above all, no dogmatism, no claims about which quantity „is" the „true" field.

2. Show to the students, that there is more than one analogy in electromagnetism.

Friedrich Herrmann

[1] F. Herrmann: *Historical burdens on physics*, 43, The field of permanent magnets

[2] F. Herrmann: *Historical burdens on physics*, 60. Inductivity

[3] A. Sommerfeld: Elektrodynamik, 4. Auflage, Akademische Verlagsgesellschaft Geest & Portig, Leipzig, 1964, Vorwort, S. VI

3.30 Poynting vector and Maxwell stress tensor

Subject:

Commonly, the Poynting vector is defined something like this:

> The Poynting vector $\vec{S}$ represents the directional energy flux of an electromagnetic field. It is calculated as the cross product of electric field strength $\vec{E}$ and magnetic field strength $\vec{H}$:
>
> $\vec{S} = \vec{E} \times \vec{H}$

Deficiencies:

Why does the energy flux density in the electromagnetic field need its own name and symbol? The meaning of a physical quantity is that it measures one and the same property at the most different physical systems. It allows to compare the systems with each other, for example the inertial behavior of the earth with that of an electron with the help of the mass. Thus one can also say: The mass of the earth is large, that of the electron is small. If one had given different names to the masses of earth and electron, the corresponding statement would be more complicated.

Furthermore, one might ask: If one gives a name of its own to the energy flux density in the electromagnetic field, would it not be consistent to give proper names to the energy flux densities in other systems as well, e.g. that in the hydraulic line of an excavator, which is calculated as $p\vec{v}$ from pressure and velocity?

Of course, the corresponding procedure is quite common in other contexts. Consider, for example, the quantity, which one sometimes calls distance, but in another context also path length, length, width, height, displacement, radius or diameter. However, these are measures which are easily accessible to our perception and which are firmly anchored in the colloquial language.

The situation is different with energy, and especially in the context of fields. One often emphasizes that the “field concept” is a difficult concept, and one introduces it in a way that gives the impression that a field is nothing more than a mathematical tool that allows to calculate the force on a mass point [1].

Actually, a field is a physical system like others, like a body, a gas or a liquid. As in any other system, also in the field the standard physical quantities have certain values: energy density (mass density), energy current density, momentum, momentum current density, charge density, entropy density and depending on the state also velocity, temperature and chemical potential.

I am afraid that the naming Poynting vector promotes the idea that with the field one is dealing with something mysterious and that one must not take the energy flow in the electromagnetic field quite seriously, or that it is something that is somehow different from a “real”energy flow.

Origin:

Formula (1) dates from the time when we were just beginning to understand how to describe energy distributions locally.

A peculiarity was that the still new quantity energy was a physical quantity of which one could not say which property it measured. Its values could only be inferred or calculated from other measurable quantities - albeit in a different way depending on the system and the state. Only in 1905, 20 years after the introduction of the Poynting vector, it became clear that energy, like any other physical quantity, measures a specific property, namely inertia and gravity.

A similar argument applies to the Maxwell stress tensor. Instead of saying that the formula found by Maxwell allows us to calculate the mechanical stresses in the field, the stresses are called Maxwellian stresses, as if they differed in anything from the normal, “real” mechanical stresses.

Disposal:

Write the formula something like this:

$$\vec{j}_E = \vec{E} \times \vec{H}$$

where $\vec{j}_E$ is the energy flux density.

But where then remains Poynting, whom one wanted to honor with the name of the vector, after all? We think it would be more suitable to connect the name Poynting not with the vector, but with the equation. This could be called Poynting-Heaviside formula. (Yes, it is true: it was found by both independently).

[1] F. Herrmann, *Historical Burdens on Physics*, 3.8 The field as a region of space with properties,

http://www.physikdidaktik.uni-karlsruhe.de/download/historical_burdens.pdf

Friedrich Herrmann

4

Thermodynamics

4.1 Mechanics versus thermodynamics

Subject:

Mechanics is the most important subject area of physics. It is the basis of physics. Thermodynamics, on the contrary, is one of several less important specialties. This is the wide-spread opinion. It can be seen when considering curricula, degree programs and text books. A typical course for a teaching degree for the secondary school comprehends 6 contact hours per week of mechanics, but only 2,5 hours of thermodynamics. The ratio is similar for the number of pages in textbooks for the secondary school and for the university.

Often it is said explicitly that mechanics has an outstanding importance. In a secondary school book we found in the context of the equation $F = m \cdot a$: "This is really the most important statement in this book; it has changed the world since 1686."

Deficiencies:

Thermodynamics is not only much shorter than mechanics in the teaching curriculum. Moreover, what is provided in the curriculum is often not fulfilled. According to the school curriculum usually it should be treated in the 11th grade, after mechanics, i.e. during the rest of the school year. However, in the turbulence of the end of the school year thermodynamics is often sacrificed.The situation is similar at the university. Often thermodynamics is taught in a one-semester course together with optics. One begins with optics but then time is running out and thermodynamics holds the short end of the stick. As a consequence many students leave the school and also the university as thermodynamic illiterates.

From today's perspective, mechanics does not deserve this preferential position, and thermodynamics does not deserve its bad reputation.

Why should just the equation $F = m \cdot a$ be so important? It is essentially Newton's second law, namely $F = dp/dt$, which expresses the conservation of momentum. But the conservation of momentum is not unique; there are similar laws for energy, angular momentum and electric charge. And the series also includes the law that makes a statement about the non-conservation of entropy: entropy can be created but not destroyed.

Origin:

We cannot express it more clearly than it was done in 1883 by Ernst Mach:

"When the french Encyclopaedists of the 18th century believed to be near to their aim to explain the whole of nature physico-mechanically, when Laplace imagines an intelligent demon which would be able to predict the course of the world in all future times from the only knowledge of the initial positions and velocities, this joyful over-estimation of the extent of the physical and mechanical insights was forgivable in the 18th Century; it is indeed a gracious, noble, edifying drama, and we can easily share the joy.

But a century later, after we have become more prudent, the vision of the Encyclopaedists appears to us as a mechanical mythology in comparable to the ancient animistic religions. Both views contain undue and fantastic exaggerations of a one-sided knowledge."

Disposal:

It is not easy, since there is a long teaching tradition. One might begin the deconstruction of mechanics by cutting back kinematics.

Friedrich Herrmann

4.2 State variables

Subject:

In connection with the introduction of the First Law of Thermodynamics it is often stressed that the internal energy is a state variable. When the entropy is introduced, one also insists that it is a state variable. Recently, the name state variable is also being used in connection with the pressure, particularly in the school book literature.

Deficiencies:

The name state variable was introduced in order to express that a physical quantity in a state has a certain value. However, this is true for all physical quantities, with only two exceptions: work and heat. If one stresses for only a few quantities that they are state variables, the impression results that being a state variable is not the normal case, but the exception. The fact that a quantity in a well-defined state has a certain value is a characteristic which one expects anyway. If one wants to emphasize something, then one should instead stress that there are two quantities, work and heat, which do not correspond to the reasonable expectation.

Origin:

It is somewhat different with internal energy, entropy and pressure.

The first formulation of the conservation of energy appeared in the First Law of Thermodynamics, which related the process variables work and heat to the internal energy. Scientists were happy to point out that the internal energy is a state variable, stressing that one of the terms is a quantity with normal properties. It appeared remarkable that the sum of two non-state variables results in a state variable.

Now to the entropy. Since the beginnings of thermodynamics it was an aim to introduce a quantitative measure for what in colloquial terms would be called heat. It went without saying that this should be a state variable. At the end of the 18th century Joseph Black introduced such a quantity. From a modern point of view, Black‘s heat is best identified with the entropy. However, since the middle of the19th century, the name heat was redefined as a so-called form of energy, i.e. as a non-state variable. Thereby, Black’s heat disappeared from physics, until it was reintroduced by Clausius with the name entropy. Since Clausius defined the entropy via the non-state variable heat, it appeared worthwhile to emphasize that the entropy is a state variable. Only much later it was recognized that the newly introduced entropy was essentially identical with the heat concept from the time of Black and Carnot [1, 2].

Pressure is often introduced via the force. A force is always exerted by a body on another. Force is clearly a concept from the time when mechanical interactions were interpreted as actions at a distance. As a result, it is natural that the student looks for a body which exerts the pressure, and one on which it is exerted. In order to put him off from this wrong expectation, one stresses that pressure is a state variable. This explanation is only needed because the pressure was introduced inappropriately from the beginning [3].

Disposal:

Small solution: One does not say that internal energy and entropy are state variables – that should be clear anyway – but one stresses that work and heat are two unusual constructions, which do not fit in the pattern of the other physical quantities.

Large solution: One does completely without the introduction of separate symbols and names for the expressions which one usually calls work and heat. As a teacher one might at first have the feeling that something important is missing. But one will soon discover that nothing is missing, and that at the same time one gets rid of some conceptual problems.

One can also confidently omit the designation “state variable” in connection with pressure. Instead of introducing the simple quantity pressure via the difficult quantity force, one introduces the pressure as an independent quantity, for instance beginning with a pressure difference: A pressure difference is the cause for a water or air flow. Then it is no longer necessary to mention that pressure is a totally normal quantity – a “state variable”. The suspicion that it would be otherwise does not arise.

[1] Callendar, H. L.: The caloric Theory of Heat and Carnot's Principle. – Proc. Phys. Soc. London 23 (1911). – S. 153: “Finally, in 1865 when [the importance of caloric] was more fully recognised, Clausius gave it the name of 'entropy', and defined it as the integral of *dQ/T*. Such a definition appeals to the mathematician only. In justice to Carnot, it should be called caloric, and defined directly by his equation..., which any schoolboy could understand. Even the mathematician would gain by thinking of a caloric as a fluid, like electricity, capable of being generated by friction or other irreversible processes.”

[2] Job, G.: Neudarstellung der Wärmelehre – Die Entropie als Wärme. – Akademische Verlagsgesellschaft, Frankfurt am Main 1972.

[3] Herrmann, F.: Einige Vorschläge zur Einführung des Drucks. – In: Praxis der Naturwissenschaften5 (1997). – S. 37

Friedrich Herrmann

4.3 Names of the ideal gas law

Subject:

The equation $p \cdot V = n \cdot R \cdot T$ is introduced under different names: gas equation, general gas equation, universal gas law, thermal equation of state of the ideal gas, and others. Since the equation relates more than two variables, one may be interested in the relationship between only two of these quantities, keeping the remaining variables constant. The corresponding relations are known under particular names. The relation between p and V is Boyle's law, the V-T relation is called Charles' law, the p-T proportionality is Amontons' law and the V-n relationship is Avogadro's law. In the French and German literature Boyle' law is called Boyle-Mariotte's law and Charles' law is called Gay-Lussac's law.

Deficiencies:

1. The importance of an equation can be emphasized by giving it a proper name. Such a name also facilitates the reference to the equation. The gas equation (let us here call it so) is important. It is valid for matter in a very large sense, provided that the corresponding substance is sufficiently diluted and/or the temperature is high enough. The equation not only applies to gases in the usual sense, as for instance the air around us, but also for the solute in a diluted solution or for the compressed plasma in the central region of the Sun. Thus, the equation is worthy of a name. It is another question wether the attributes "general" or "universal" are appropriate, since such a classification can hardly be topped.

2. It is a nice custom to name equations after an important scientist. However, as the gas equation shows, this can also be overdone. In our case six researchers are honored by means of one single equation. The problem of baptizing an equation with the name of a scientist is known from street names. Someone may come to unexpected honors since a small alleyway that carried his name transformed later into a main artery. On the contrary, there are great scientist who never became the patron of an important equation, as for instance Leibniz or Descartes. Still others are honored for something that was relatively unimportant in their work, as for instance Huygens for the elementary waves or Faraday for the somewhat puny Faraday effect or the curious Faraday cup.

3. Let us come back to the gas equation. It is equivalent to various other relations that seemingly state something rather different from the gas equation in its usual form:

(a) $E(V) - E(V_0) = 0$ for $T = \text{const}$,

in words: At fixed temperature the energy of a gas is independent of the volume.

(b) $S(V) - S(V_0) = n \cdot R \cdot \ln\frac{V}{V_0}$, for $T = \text{const}$,

in words: At fixed temperature the entropy depends logarithmically on the volume.

(c) $\mu(p) - \mu(p_0) = R \cdot T \cdot \ln\frac{p}{p_0}$, for $T = \text{const}$,

in words: At fixed temperature the chemical potential depends logarithmically on the pressure. (From this equation one easily obtains the law of mass action and the barometric formula.)

All of these three equations can be derived without any further physical input from the "gas equation". Therefore, each of them could also be called "gas equation", what is not done.

4. A gas is not fully described by the gas equation or "thermal equation of state". The thermal equation of state is just one of several equations of state that are needed to completely characterize a particular gas. So it does not describe the caloric properties of an ideal gas: How does the temperature depend on the heat (entropy) content of the gas? The answer to this question is given by the "caloric equation of state". The effects that it describes are as striking as those described by the thermal equation of state. Traditionally, at school it is considered less important, with the result that many interesting processes are disregarded in the classroom: the isentropic expansion in the steam engine and the internal combustion engine, or the decrease of the temperature with height.

Origin:

1. In the usual treatment of the gas equation and various proportionalities which it contains, one can recognize the various contributions of the various epochs of its genesis. One also sees that the view from different countries is different.

2. The thermal equation of state, that contains the variables p, V and T, which are easy to measure, is overrated since the quantities entropy and chemical potential, which for many processes are more important, and, by the way are also easy to measure, never found wide acceptance.

Disposal:

1. Introduce names for equations parsimoniously. In the case of the gas equation we propose not to give own names to the partial proportionalities. We recommend particular caution in the award of predicates like "general", "universal", "fundamental" etc.

2. If the logarithm is available, discuss the volume dependency of the entropy and the pressure dependency of the chemical potential. Treat in any case the "caloric" properties of gases, in particular the relation between temperature and volume at constant entropy, since it allows for an understanding of the working principle of the heat engine and the temperature stratification of the atmosphere.

Friedrich Herrmann

4.4 Preliminary temperature scales

Subject:

Usually the temperature scale is introduced by referring to the thermal expansion, in particular that of gases, and temperature is defined only subsequently, if at all, as an "absolute" quantity T, i.e. a quantity that is independent of any thermometric substance, by means of the postulation that the efficiency η of a Carnot process with the working temperatures T_+ and T_- is given by $\eta = (T_+ - T_-)/T_+$.

Deficiencies:

Considering a common behavior of gases of low density, so that a unique thermometric scale θ can be defined for them, and forgetting all the other thermometric substances, we remain with two temperature quantities, θ and T, whose definition, handling and relationship has to be discussed. The result is $\theta \sim T$ and therefore $\theta \equiv T$, if for only one point –for instance the triple point of water– $\theta = T$ is required. The derivation is not difficult, but it is rarely given.

It is easier to abstain completely from provisional temperature scales and Carnot processes and instead metricize the entropy S directly and define the temperature T departing from the entropy flow I_S and the energy flow P, and using $P = T \cdot I_S$. It is easy to show that the quantity T introduced in this way has all the familiar properties of a temperature and can be measured with a conventional thermometer. For the calibration one can take advantage of the simple fact that if the internal energy only depends on T (as is true for gases of low density), the pressure p at constant volume must be proportional to T, so that the temperature can be determined by measuring p. Moreover, the physics of simple heat engines and heat pumps just falls into one's lap. Two or three lines and the four basic arithmetic operations are sufficient, whereas normally work and heat diagrams and the differential and integral calculus of functions with several variables are employed.

Origin:

The textbooks follow strictly the historical development. The fact that liquid expansion thermometers are still in use makes it appear natural that we introduce the concept of temperature via thermal expansion. Since the gas equation is an important teaching objective, the treatment of the gas-thermometric temperature scale seems obvious. Otherwise, according to the general conviction entropy is beyond the school horizon and thus outside of the field of view of the pedagogues.

Disposal:

It can only be successful if we abandon the prejudices against the entropy which have been nourished for one and a half century and belong to our physical educational background, and that put it down as a complicated state variable in an abstract calculus and that deny it without reason any property that can be grasped with our senses.

Georg Job

4.5 Thermal expansion of liquids and solids

Subject:

The thermal expansion of liquids and solids

Deficiencies:

For temperature changes of 10 °C the effect is of the order of per mil. There are many other effects of this order of magnitude. In general we cannot afford to treat so small effects when teaching physics to beginners.

An argument in favor of the subject might be that there are phenomena in our everyday life that can be explained by the thermal expansion of liquid or solid bodies. An example is the Mercury thermometer. There are, however, other types of thermometers, that are just as important and for which we do not spend any teaching time.

Another example that is mentioned when teaching the subject is the expansion of railway rails and bridges. We believe that this subject is arguable. When explaining the expansion of bridges one should also explain why most of the other objects do not expand upon heating: houses, streets or even the whole Earth. And when citing the railway rails one should explain why in former times there was a gap at the rail bond and nowadays there is non.

There is another effect that sometimes is confused with the thermal expansion at constant pressure: the change of pressure at constant volume. This effect is big and impressive, in contrast to the thermal expansion. Both effects – volume change at constant pressure and pressure change at constant volume – are governed by two independent coefficients. The pressure change at constant volume in liquids and solids is normally not treated in the beginner's course. The reason seems to be that even this big effect is not important enough. Every subject competes with many other subjects. And there are so many that are important enough, but we do not have the time to discuss them.

Origin:

The reason why thermal expansion of liquids and solids occupies so much teaching time is not its importance in applications. It is the old mercury temperature scale. Before the thermodynamic temperature scale was introduced into physics, the definition of temperature depended on the thermal expansion of mercury. But this argument is not valid anymore.

Disposal:

Devote less time to the subject. There would be no harm if it were completely omitted.

Friedrich Herrmann

4.6 Amount of heat and heat capacity

Subject:

The process quantity Q that appears in the first law of thermodynamics is called amount of heat, heat energy or simply heat. The heat capacity C is defined as the quotient of the heat ΔQ which is supplied to a system and the resulting temperature increase ΔT:

$$C := \Delta Q / \Delta T \,. \qquad (1)$$

Usually the specific heat capacity c is used which one obtains by dividing C by the mass. For our purpose it is sufficient to consider C.

Deficiencies:

The concept “capacity” has to do with storage. The problem becomes obvious in the name of the “quantity” which is to be stored: the heat Q. When calling Q amount of heat or heat energy, one suggests that Q has for a system in a given state a well-defined value, and that this value refers to a given region of space, or in other words: that Q is an extensive or “substance-like” quantity. Actually Q does not have these properties. Q is not a physical quantity in the usual meaning of the term, but $dQ = TdS$ is a so-called differential form. For that reason it is impossible to use it as a measure of the heat content of a system. While it is acceptable to say that one supplies the amount of heat dQ to a system, it is not correct to say that thereby the heat content of the system changes by dQ. It sounds like sophistry but it is not.

The problem is simply an inconvenient naming. The name heat or amount of heat suggests that the corresponding quantity is an extensive quantity. But Q or dQ is not an extensive quantity. It would be better not to give to this differential form a proper name nor a proper symbol. Then one would not raise expectations that later cannot be satisfied.

Now, if a heat content cannot be defined by Q, then the quantity defined by equation (1) cannot be interpreted as a heat capacity.

Who has ever taught thermodynamics knows how difficult it is to make the students understand that Q is not a state variable. The corresponding efforts are foiled by the inconvenient name heat capacity.

Origin:

Both terms “heat” and “heat capacity” came into use in the 18. century, i.e. an epoch when the process quantity Q did not yet exist. The term heat was used for a substance-like state variable which measured what in colloquial terms would be called heat or amount of heat. When the energy came into being at the middle of the 19th century, the old state quantity heat was robbed of its name, and henceforth the name heat was attributed to a so-called energy form. Since the new heat had the unpleasant property of not having a value in a given state, it was euphemistically called a process quantity.

Disposal:

When doing thermodynamics operate from the very beginning with entropy. It can be introduced as that quantity which measures what in colloquial terms is called heat. It can always be said how much of it is contained or stored within a system and it makes sense to define a corresponding capacity dS/dT. The entropy capacity is related to the established “heat capacity” in a simple way:

$$C = T \cdot dS/dT \,.$$

(Incidentally the entropy capacity spooks in some solid state physics text books under the name of Sommerfeld constant [1].)

[1] Kubo, R. und Nagamiya, T.: Solid State Physics. – McGraw-Hill Book Co., New York 1969. – S. 94

Friedrich Herrmann

4.7 Heat transfer

Subject:

From a school book:

"The transmission of energy in the form of heat can take place as heat conduction, convection or heat radiation."

From a scientific encyclopedia, under the keyword *heat transfer*:

"The transfer of heat can take place in three different ways: By heat conduction, where heat flows through a solid medium or a stagnant fluid; by convection, where the heat is transported through the movement of the medium (usually a fluid); by radiation, where the heat is transmitted in the form of electromagnetic waves."

Deficiencies:

The sentence "In our garden there are trees, useful plants, vegetables and weeds" is not incorrect, but something is disturbing. It suggests that the four categories of plants that are mentioned represent a classification of the plants of the garden, which it is not. There are plants that correspond to more than one category, for instance carrots or cherry trees.

Our citations about heat transfer are of the same kind. They suggest that any transfer corresponds to one of the categories conduction, convection or radiation, or at least that it is possible to say to what extent it corresponds to one or the other type of transport. However, there are heat transfer phenomena that cannot be classified according to this scheme.

Before considering some examples, let us characterize the three categories in some more detail:

1. Heat conduction: Heat and entropy flows through a material; the material does not move. The transport is dissipative, i.e. new entropy is produced. The "driving force" for the process is a temperature gradient.

2. Convection: Heat and entropy is supplied to a material (e.g. air or water). The material moves and takes the entropy with it. No temperature gradient is needed for the transport. However, in order to move the material, there must be another driving force, for instance a pressure gradient.

3. Heat radiation: The heat is transported by electromagnetic radiation.

Now it may be seen more easily that these categories do not define a classification. The first and the second definition (conduction and convection) are characterized by the nature of the driving force. They say nothing about the type of carrier particles. Indeed, the first category comprises heat conduction in non-metals (the carriers are phonons), in metals (the carriers are electrons) and in gases (the carrier particles are molecules).

The third category on the contrary is only characterized by the carrier particles, namely photons. An additional problem with this third category is that not every energy transport with photons can be considered heat radiation. Microwaves of a single wavelength do not transport entropy. Thus there is no reason to call such a transport a heat transfer [1].

An example of a heat transport that does not fit into this scheme is the heat flow within the sun from the reaction zone to the surface. (Only the outer 10% of the way from the reaction zone to the surface is convective. We are not interested in it for the moment.) The carriers of the energy and the entropy in the sun are photons. Does the transport therefore belong to the third category? Each photon runs only a short distance and is then absorbed, and a new photon is emitted, which runs in any arbitrary direction. Thus the radiation goes forth and back, right and left, up and down; in other words: it is a diffusive process, and thus similar to normal heat conduction, but with photons instead of phonons. Like normal heat conduction it occurs only if there is a temperature gradient. (In the sun this is 30 K/km on the average.) We conclude that this process does not only belong to the third category, but also the first.

Another example: Entropy is transferred from the surface of the Earth to the upper side of the troposphere, from where it is irradiated into space. A small part of this transport (8 %) takes place in a way that is similar to the heat transport process within the sun: The Earth emits electromagnetic radiation. This is absorbed and reemitted again and again. Thus, this process too belongs to both the first and the third category.

There are other heat transport processes which cause difficulties when trying to attribute them to one of the three categories: A beam of thermalized atoms, electrons or ions, or a flow of thermal neutrinos in a collapsing star.

In this context we may also ask the question: If it is supposed that there are three kinds of heat currents, shouldn't we also expect that there are several kinds of electric currents and several types of momentum currents etc.? And indeed, we can introduce categories for other currents. However, we then will encounter the same kind of difficulties as with the heat currents. So what kinds of electric currents do exist? There are currents realized with electrons in metallic conductors, with ions in a solution, there are free electron beams, currents realized with various carriers in plasmas, moving conductors, supercurrents of the first and the second type. Obviously, one would not claim apodictically: "There are three (or four or ten) kinds of the transport of electric charge."

Origin:

The rule came into being at a time when three kinds of heat transport were known and it was concluded somewhat carelessly, that there were no others, and that these defined a classification. This happened when nothing was known yet about the interior of the sun, about the heat transport within the Earth's atmosphere, nothing about electromagnetic fields, photons, plasmas, and nothing about neutrinos.

It cannot be excluded that also some number mystery played a role: "All good things come in threes."

Disposal:

If one is really willing, one might try a more sustainable classification. One will then note that several classifications can be made according the which criteria are chosen. The criteria could be:

– Which are the carrier particles?

– Is the transport dissipative or not?

– Which is the driving force: a temperature gradient, another gradient, or no gradient at all?

One would notice however, that not much is gained by any of these classifications. So, our recommendation is, (and not only here): abstain from classifying.

But what about those processes that initially were called conduction, convection and radiation? We certainly want to distinguish between them, just as we want to distinguish carrots from stinging nettles. But in both cases we don't need a classification scheme.

[1] *F. Herrmann and P. Würfel:* Thermal radiation, article 4.32

Friedrich Herrmann

4.8 The equivalence of heat and work

Subject:

Heat is disordered energy according to some [1], the kinetic energy of the unordered molecular movement according to others [2]. To still others it is the kinetic and potential energy of the thermal molecular motion [3], the energy which can be added to an object by thermal contact [4], a short name for the expression $U - W$ [5], the bound energy TS [6], the integral $\int TdS$ [7], or a questionable and superfluous concept [8]. What is it really?

Deficiencies:

Clausius himself used two heat measures: the heat H contained in an object, which he imagined as kinetic energy of the molecules, and the "supplied heat" Q, where $Q = H$ is only valid in exceptional cases. Under the above cited examples we easily recognize the descendants of these two parents. The diversity of opinions is an expression of the unpleasant circumstance that no energetic quantity exists that simultaneously represents all the desirable aspects of the concept of heat. As with a too-short blanket, one is obliged to do without one or the other property. Depending on what one chooses to stress, the compromise will be different. From the fact that in spite of this ambiguity the mathematical treatment gives the same results, one can conclude that the equivalence postulated by Clausius is meaningless for thermodynamic calculations. Then what is it good for?

Origin:

The question is as old as physics. The answer given by R. Clausius around 1850 in his first law, in which he postulated the equivalence of heat and work, is essentially still considered valid today, but it is obviously ambiguous.

Disposal:

If we do without this postulate, we win a new freedom. We do not need it in order to formulate the conservation of energy. We also don't need it to define what heat is. It is easy to give an operational ("fundamental") definition of the heat concept. Such a procedure is usually used in physics only for the definition of some basic quantities, such as length, time and mass. One specifies the unit and how to determine equality to the unit and multiples of it. However, one can employ this procedure, which makes a correspondence between a given concept and a physical quantity, in many other cases, for instance for the definition of energy, momentum, angular momentum, charge, amount of material, entropy, and for the metrization for concepts like the amount of heat, amount of data, disorder or randomness. The most surprising result from proceeding in this way is that a concept of an amount of heat that is unbiased by scientific interpretation does not result in an energetic quantity, but in Clausius's entropy S [9]. This effortless access to the most important thermodynamic quantity (apart from the temperature) permits a far-reaching house-cleaning of thermodynamics. Concepts such as enthalpy, free energy, energy degradation, process quantity and state function can be disposed of at the same time. The fact that a small error can have such far-reaching consequences, not in the scientific calculus but in its semantics, should warn theoreticians, whose attention is focused mainly on the consistency of the calculations, and alarm educationists, who have to deal with the consequences.

[1] F.J. Dyson:"What is heat?" Scientific American 1954, 191 (No. 3), S. 58 - 63.

[2] R.W. Pohl: "Mechanik, Akustik, Wärmelehre"; Springer: Berlin 1962, S. 248.

[3] C. Gerthsen, O.H. Kneser, H. Vogel: "Physik"; Springer: Berlin 1986, S. 193 - 197.

[4] C. Kittel: "Physik der Wärme"; Wiley & Sons: Frankfurt 1973, S. 133.

[5] M. Born: Physikal. Zeitschr. 1921, 22, S. 218 - 286.

[6] H.H. Steinour: "Heat and Entropy"; J. Chem. Educ. 1948, 25, S. 15 - 20.

[7] G. Falk, W. Ruppel: "Energie und Entropie"; Springer: Berlin 1976, S. 92.

[8] G.M. Barrow: "Thermodynamics…"; J. Chem. Educ. 1988, 65, S. 122 - 125.

[9] The following assumptions together with the choice of a heat unit are sufficient for its unambiguous metrization:

1) each object contains heat, whose amount cannot decrease, if it is thermally insulated.

2) objects of the same kind and in the same state contain equal amounts of heat.

3) the heat content of a composed object is equal to the sum of the heat contents of its parts.

Georg Job

4.9 Thermal energy

Subject:

From a school book:

"The *thermal energy* is a part of the internal energy and essentially determined by the temperature. Since in many cases one can presuppose the constancy of the other components, one often only considers the thermal energy…

Heat tells us how much thermal energy is transferred from one system to another.…

The following relation holds between transferred heat and energy change:

$Q = \Delta E_{\text{thermal}}$."

From another school book:

"The potential and the kinetic energy of the particles taken together is called thermal energy."

From a third school book:

"The total energy of a thermodynamic system, which consists of *thermal energy* (*potential* and *kinetic energy* of the particles), of *chemical energy* and *nuclear energy* is the *internal energy U*."

Deficiencies:

The intention of these definitions of thermal energy is clear: The authors of the statements try to define a quantity which measures the "heat content" of a system and which has the following properties:

1. It should be a state variable, i.e. it should have a well-defined value for a system in a given state.

2. It should be an energetic quantity, i.e. a quantity that is measured in Joule.

3. It should be a part of the internal energy. Another part would be the chemical energy.

4. Differences of it should be equal to the process quantity *Q,* which in physics is called heat.

The problem is that a quantity that meets these requirements does not exist and cannot be defined. It is not possible to distinguish the potential and kinetic energy of particles from a part which might be called chemical energy. Any temperature increase is related to electronic excitations, to oscillations, to excitations of the spin system, to the dissociation of molecules, to a rearrangement of atoms, i.e. chemical reactions, and finally to nuclear reactions. There is no possibility to decompose the energy that is engaged in these processes in an unambiguous way into a thermal and a chemical component. If such a decomposition were possible, it would manifest itself in the fact that one summand (the thermal energy) would depend only on temperature and not on the chemical potential and another summand only on the chemical potential and not on temperature. But such a decomposition is not possible.

Origin:

Physics, chemistry and technical thermodynamics need a measure for the heat content of a system. Common sense suggests that it should be possible to define it, since we intuitively operate successfully with such a quantity. However, when trying to define a measure for heat in the 19th century, a mistake was made: It was supposed that such a quantity should be an energetic quantity. However, a definition of an energetic quantity with the desired properties could not work. As a result several surrogates appeared, each of which satisfies some of the requirements and others not. The quantity *Q*, which was called heat, is one of them. The problem is that *Q* is not a physical quantity in the usual sense of the word. One says that it is a "process quantity" since it makes no sense to ask for its values for a given system in a given state. Chemists prefer to manage with another "surrogate" quantity, the enthalpy. This quantity behaves like a heat content, but only as long as one restricts to processes at constant pressure – for the physicist an unacceptable restriction.

None of the quantities *Q* and *H* meets the justified expectation towards a measure of a heat content. So, why not define a quantity that better suits to our needs, the thermal energy?

It is interesting that the concept "thermal energy" can only be found in school books, but not in University texts. Do we have to reproach to the school text book authors for inventing untenable concepts, due to their ignorance of thermodynamics? Yes and no. Yes, because the definition does not work. No, because they are not to blame for the fact that thermodynamics is so unfamiliar and so unpopular.

It is the University that is to blame. Here, what students learn about thermal phenomena: Relations between four quantities that change their values simultaneously, interlaced partial derivatives, changes of variables, unintuitive quantities like enthalpy, free energy, and Gibbs' free energy are the requisites of the chamber of horror. For the simple explanation of the compression of the gas in a Diesel engine the so-called adiabatic index is used, which is defined as the quotient of two partial derivatives, which are distinguished by the fact that in one of them one variable is kept constant in the other another variable.

It is not to expect that the students get a non-tensed relation to thermal phenomena in this way. But how can he or she, later as a teacher, present thermal facts to beginners? It is understandable that the school teachers and school book authors try to construct a thermodynamics that does not offend common sense.

Disposal:

It is much simpler than one might believe. It is sufficient to abstain from demanding that a measure of heat must be an energetic quantity. All the difficulties disappear when introducing entropy as the measure for a heat content.

Friedrich Herrmann
Georg Job

4.10 Internal energy and heat

Subject:

If heat is supplied to a body, then the body will contain more heat. If the body delivers heat, then at the end it has less. A person who is not educated in physics will surely not object to these statements. However, physics teaches us that they are incorrect: One can supply heat to a body, but thereafter it has none, and although it does not possess heat, one can extract heat from it. It looks like magic. The top hat is empty, but out of it comes a rabbit. Physics tells us that supplying or extracting heat does not change the heat content of a system, it changes the internal energy or enthalpy, depending on how the heat is supplied. The fact that energy is not called heat as soon as it arrives in the body is more than just a convention. There simply is no means to tell how much heat is contained in a body. In physics text books, this irritating circumstance is expressed in different ways. Some authors express it courageously [1]. Others risk doubtful justifications by maintaining that the internal energy can be divided in fractions, which they themselves would be unable to quantify [2], [3] (see also [4]). Sometimes heat and internal energy are simply taken to be identical [5].

Deficiencies:

I cannot imagine that even a single pupil will understand why it is incorrect to say that the heat supplied to a body remains inside the body. Most of our university students also would be unable to give an explanation. The statement appears to the student either only as sophistry, or it is memorized together with the numerous topics that one does not understand, and does not necessarily need to understand.

Origin:

For the description of the heat supply to a body one would need a quantitative measure of heat. The “heat” of the physicist as a “process variable” [6] is as poorly suited for this as the internal energy or the enthalpy so beloved by chemists. See also [7], [8].

Disposal:

It is particularly simple. One describes the process with the entropy. Entropy corresponds exactly to a non-physicist‘s idea of heat. If one heats something up, one supplies it with entropy, and after the entropy is supplied, the entropy is in it. It is easy to give a value for how much entropy is within a body, and still easier to quantify how much the entropy changes when warming the body up [9].

[1] Galileo 9 (Oldenbourg 2000) p. 98: “Warning! Differentiate very carefully between heat, internal energy and temperature: An object does not possess heat, but internal energy!”

[2] Spektrum Physik (Schroedel Velag Hannover 2000) p. 17: Under the heading "the portions of the internal energy" are specified: the kinetic energy of the particles; the energy, which is in the co-operation of the particles; chemical energy and nuclear energy.

[3] Galileo 9 (Oldenbourg 2000) p. 93: “The energy of an object, which is not to be described as mechanical energy (potential or kinetic energy), one calls internal energy E_i. The atomic energy, the chemical and the biological energy all belong to the internal energy. A substantial portion is also the energy which is connected with the temperature of the object.”

[4] *G. Job*, Energieformen in Altlasten der Physik, Aulis Verlag Deubner Köln, 2002, p. 11.

[5] Metzler-Physik (Metzlersche Verlagsbuchhandlung Stuttgart 1988) p. 60: “In all of these cases the bodies are performing frictional work; thereby a part of this mechanical energy is transformed into an energy form that cannot be transformed back into mechanical energy, but is given away as heat energy or internal energy to the environment inside or outside of the system.”

[6] *F. Herrmann*: State Variables, article 4.2

[7] *G. Job*: The Equivalence of Heat and Work, article 4.8

[8] *G. Job*: Entropy, article 4.13

[9] *F. Herrmann*: Measuring entropy, article 4.14

Friedrich Herrmann

4.11 Available energy

Subject:

In publications about the energy economy of a country by public authorities and universities one often finds so-called energy flow charts [1,2,3]. They indicate the energy balance of a national economy. The title may be something like: “Estimated U.S. Energy use in 2008”. They show with which energy carriers the energy enters the system, i.e. the national economy, which part is transformed into which other energy forms and in which forms the energy leaves the system. For the outgoing flow the distinction is made between final, available or useful energy on one hand, and lost or rejected energy on the other.

Deficiencies:

The following impression arises: for the applications of the final user energy is needed in a certain form. Hence it has to be transformed, and by doing so part of it is lost. One tries to keep the transformation and transportation losses as low as possible, but for physical reasons a considerable loss is unavoidable. Only when arriving at its final destination, i.e. at the end user, the energy can be employed for what it is really needed.

This view does not exactly meet the point. This can be seen when considering that every energy loss is due to the production of entropy. This entropy must be eliminated, i.e. transferred to the environment. For that purpose energy is needed and this energy is lost. The flow of the lost energy P_L is proportional to the flow of the entropy I_S that has to be disposed of:

$P_L = T_0 \cdot I_S$

Here, T_0 is the absolute temperature of the environment, i.e. of the system that absorbs the entropy.

From this consideration two conclusions can be drawn:

1. From a physical point of view the losses are not unavoidable. Any process can also be carried out in a reversible way. It may be impossible for technical or economical reasons, but it is not physics that forbids them. Even the “transformation” of the chemical energy of Carbon (+ Oxygen) into electric energy, where one usually holds the Carnot factor responsible for the low efficiency, can in principle be carried out reversibly, for example in a ideal fuel cell. Thus, in the energy flow chart, already the incoming energy could reasonably be called useful energy.

2. After arriving at the so called end user all the energy is eventually “used” to produce entropy, i.e. is wasted; all of the energy that has been sold to the user as available energy ends as lost energy. By the way: Also for the final user holds, that every process in which he is interested, can be carried out reversibly.

We do not claim that something is wrong with the flowcharts, neither that they are not useful. We only believe that they convey a false message. It is not correct, that only a fraction of the primary energy is “really” useful. After a series of steps all of the primary energy ends in entropy production, and the flowcharts mentioned above tell only half of the story.

Origin:

Why do energy flowcharts stop at a certain point? Why do they not show that all of the “usable energy” eventually ends up in the thermal depository? Because they are issued by institutions that have particular interests. For the energy industry the picture ends at the fare stage, the place where they ask for money. They are concerned with the losses before this point. They do not care about what the client is doing with the energy which they have sold him.

Disposal:

Above all, we clarify that the whole of the primary energy is used for entropy production. Since there are no physical limits to reduce the entropy production there is no physical limit for energy saving. We discuss the technical problems that arise, when trying to approach this goal. In this way students learn much Physics and also Chemistry.

[1] https://str.llnl.gov/Sep09/simon.html

[2] http://www.energyliteracy.com/?p=310

[3] http://blog.everydayscientist.com/?p=1773

Friedrich Herrmann

4.12 Tendency to the energy minimum

Subject:

The common reason given as a cause of a process is that the system reaches a state of minimum energy as a result of this process:

– a pendulum comes to rest at its low point
– a floating board tilts on its side
– a soap bubble forms in a spherical shape
– a sponge sucks up water
– a quantity of electric charge distributes on a conductor
– excited gas atoms emit photons
– positive and negative ions arrange themselves in a crystal lattice
– heavy nuclei decay.

Deficiencies:

Without saying it explicitly, all of these statements assume that each system aims at a state of minimum energy and proceeds to this state, provided it is not hindered by some circumstance. Formulated this way, however, the statement doesn't make sense. If one system reaches a state of minimum energy, then the complementary system, the environment, must reach an energy maximum due to the conservation of energy. The same argument applied to the environment would yield the opposite result. Thus the above assumption cannot be valid generally. So for which system is it valid? The answer comes from thermodynamics. The system must, as W. Gibbs expressed it in 1873, be closed for everything except the energy necessary to keep the entropy constant. The entropy S_p produced by processes occurring within the system appears only in the environment, and with it the energy TS_p coming from the system, where T is the temperature of the environment. Since S_p and T are always positive, the system always loses energy, since any other energy exchange that could compensate the losses is forbidden. Seen in this way, the tendency to the minimum energy is nothing more than a consequence of the entropy principle, applied to a particular class of systems.

Origin:

In mechanics we ignore the thermal properties of things. Levers, pulleys, springs, blocks and ropes are considered objects that cannot be heated, i.e. whose temperature and entropy cannot change. In fact we are tacitly ascribing the entropy created by friction to the environment. Under these conditions, we are allowed to speak of the tendency to an energy minimum. The same applies to systems in many other parts of physics – hydraulics, electricity, atomic and solid state physics and so on. Because we don't mention the production of entropy as the cause for these processes, we get the impression of an independent natural principle.

Disposal:

We can talk about entropy production in systems explicitly. Like so often, our strained relationship to entropy misleads us to questionable surrogates. The fundamental evil, which as a consequence has endless difficulties and opposes itself to any attempt to remedy, is the dogma of the heat as a special form of energy, which for one and a half centuries has been affectionately cared for, and which is anchored in the first law of thermodynamics. Only if we are ready for a revision can a lasting improvement be expected.

Georg Job

4.13 Entropy

Subject:

"Entropy" is the name of a quantity that is introduced in classical thermodynamics as an abstract function, defined by in integral. This approach makes the quantity so aloof that it costs quite an effort even to the specialist to deal with it. Its interpretation as a measure of disorder is an approach that is favored by chemists, in order get at least a rough understanding of its meaning.

Deficiencies:

It is an advantage of the chemist's approach that entropy can be qualitatively seized, but this is not enough to satisfy the standards of the physicist. For a physicist a quantity is defined only if he knows a procedure to determine its values. Another flaw is that when using the disorder interpretation is seems that no simple macroscopic property corresponds to entropy.

Origin:

In the first half of the 19th century it became clear that the conservation of heat, as supposed by Carnot and others, was untenable. This brought in 1850 Clausius to try a restructuring of thermodynamics by supposing that heat and work can be transformed one into the other. In the scope of his work he constructed the quantity *S*, in order to describe the limitations of this mutual transformation.

Disposal:

In 1911 in a presidential address to the Physical Society of London its then president H. L. Callendar [1] pointed out that *S* is nothing else than a complicated reconstruction of the quantity that had been called heat by Carnot, the only difference being that now heat can be produced, but, as before, not destroyed. This insight arrived half a century too late to rectify the erroneous itinerary. One could conclude, however, that the quantity *S* not only has the same obvious intuitive meaning as the old heat, but it also can be quantified in in the same simple way. Thereby the formalistic ghost *S* of classical thermodynamics could reduce to a concept that can easily be handled by a pupil of the elementary school, and the arsenal of now superfluous mathematical tool could be disposed of. This expectation is confirmed by great amount of experience at many schools [2]. In the role of heat *S* becomes, even under the featureless name of entropy, a quantity that is no more demanding than the concepts of length, time or mass. The fact that it appears in another raiment in information theory, statistical physics and the atomistic ideas of the chemists does not hinder it to appear in macrophysics in the role of heat.

[1] H. L. CALLENDAR: Proc. Phys. Society of London 23 (1911) 153. Callendar also proposes to abbreviate the legal unit J/K by Carnot.

[2] The Karlsruhe Physics Project, see for example

http://www.physikdidaktik.uni-karlsruhe.de/

Georg Job

4.14 Measuring entropy

Subject:

Often entropy is introduced in such way that the impression results that values can be obtained only through complicated mathematical calculations.

Deficiencies:

Entropy and temperature are the most important quantities of thermodynamics. Entropy is the energy-conjugated extensive quantity of the intensive quantity temperature. Between entropy and temperature there is the same relationship as between electric charge and electric potential, or between momentum and velocity. Thus it is reasonable to expect that in the teaching of thermodynamics entropy and entropy currents should play an important role, just as electric charge and electric currents in electricity or momentum and force (= momentum current) in mechanics. The usual introduction of entropy does not cope with this expectation.

When mentioned for the first time, it is usually stressed that entropy is a state function [1]. But why is it claimed that it is a *function*? In the first place entropy is a physical quantity or in mathematical terms, a variable. It is a function only if its dependence on other physical quantities is given. And depending on which other quantities we choose, this function is different.

And why is it stressed that entropy is a *state* function or variable? Almost all physical quantities are state variables. This fact, however, is so evident that it is not considered to be worth mentioning. Only because the traditional introduction of entropy does not allow for an intuitive idea of the quantity, one clings to this property, although it is not at all distinctive.

The most important deficiency when introducing entropy is that no measuring procedure is presented. The complicated introduction makes us believe that a measurement is difficult, if not impossible.

Actually, entropy is one of those quantities that are the easiest to measure. Entropy values can be determined with good precision with only the aid of kitchen equipment.

Origin:

See [2]

Disposal:

Of course, we do not want to dispose entropy and its measurement, but the prejudice that entropy is hard to measure.

How can entropy be measured? Let us first define the task more exactly: Determine the entropy difference of 5 liters of water at 60 °C and at 20 °C.

We begin with the water at 20 °C and heat it with an immersion heater until its temperature is 60 °C. When doing so we stir the water and measure the temperature as a function of time. The energy flow or power P from the heater into the water is known.

From $dE = TdS$ follows

$$dS = \frac{dE}{T} = \frac{Pdt}{T} \ .$$

We thus get the value of a small increase of the entropy dS as the quotient of the delivered energy $dE = Pdt$ and the absolute temperature T. Since the temperature changes as we are heating, we have to integrate or to sum from T_1 to T_2 in order to get the total entropy:

$$\Delta S = P\int_{T_1}^{T_2}\frac{dt}{T} \approx P\sum_i \frac{\Delta t_i}{T_i}$$

As long as the temperature changes are not too great compared to the average absolute temperature $\overline{T}$ we get approximately

$$\Delta S \approx \frac{P\Delta t}{\overline{T}} \ ,$$

or in words: The increase of the entropy is equal to the power of the immersion heater multiplied by the heating time and divided by the absolute temperature.

[1] Gerthsen – Kneser – Vogel: Physik. – Springer-Verlag, Berlin, 1977. – S. 183

[2] *G. Job*: Entropy, article 4.13

Friedrich Herrmann

4.15 What actually is energy? What actually is entropy?

Subject:

Last July in a symposium of the WCPE conference (World Conference on Physics Education) about the introduction of energy in the classroom speakers reiterated that energy is an abstract concept. By saying so they were in good company. We also know from Feynman [1]: “It is important to realize that in physics today, we have no knowledge what energy is.”

It is not different with entropy. The great John von Neumann explains [2]: “No one knows what entropy really is, so that in a debate one is always at an advantage.” Or the mathematician Harro Heuser [3]: “The concept of entropy belongs to the most occult concepts of physics .... ”

Deficiencies:

What actually is energy? What actually is entropy?

1. The questions are ill-posed, the “actually” should get out. And the above-cited remarks are unnecessarily fatalistic. Instead of issuing warnings or intimidations, our first response to the questions should be: Both are physical quantities, and thus a measure of something. Of course, this response gives rise to a new question: For what are they a measure? But this question sounds by far not as transcendent as the original question, and it can be answered.

2. Usually, one only asks for these two quantities, but never it is asked what is “actually” mass, electric charge, temperature or pressure. These questions are just as interesting, and it is just as hard or easy to answer them.

3. A well-asked question defines an ensemble of responses, one of which is the correct one. The question: “Who has obtained grade A in the test?” specifies that the response is one of the names of the students. There are also less well-asked questions and there are bad questions. A bad question does not define a set of answers, and that is why it is almost impossible to satisfy the questioner. Parents can tell you a thing or two about it. These are above all the questions that begin with “why”. You answer them and there comes the next “why”. The same class of questions are those that begin with: “What actually is…?” With what kind of answer might the questioner be satisfied?

4. Let us try to give an answer.

What is energy? The same as mass. The corresponding formula is cited more often than any other physical formula, but more than a hundred years after its discovery the message has not arrived. Of course, one could go on asking: “And what is mass?” But quite rightly nobody asks this question. One may object that Einstein’s answer is not convenient in the context of classical physics, since the theory of relativity manifests itself only at high velocities. This objection is not relevant. The properties of mass-energy are inertia and weight, and they manifest themselves everywhere. The fact that we cannot distinguish between a full and an empty battery by weighing it, has nothing to do with a velocity that is not relativistic enough; it is due to the fact that the precision of our scales is not good enough.

And what is entropy? It is what in colloquial terms we would call heat. Who opines that this answer is too simple may go on asking: But what is the microscopic manifestation of entropy? and so on. But he should not say that we do not know what entropy “actually” is.

Origin:

There is a misunderstanding that came about historically: Energy is a “something”, a kind of substance, that exists in the real world, and whose nature is difficult to fathom.

One does not ask the corresponding question related to the electric charge, because one believes to know what it “actually” is. But this conviction is based on a misunderstanding: One often confounds the physical quantity “electric charge” with the physical system “electron”.

Disposal:

Introduce both quantities, energy and entropy, by means of a model. Imagine each of them to be a kind of fluid. Energy and entropy are measures for the amount of these imagined fluids. One of these fluids represents inertia and weight, the other represents heat. Such an introduction is easy to understand and it is a solid basis for a serious description of the physical world.

[1] *R. P. Feynman, R. B. Leighton, M. Sands*: The Feynman Lectures on Physics (1964), Volume I, 4-1

[2] *M. Tribus, E. C. McIrvine*: Energy and Information, Sci. Am. 224, Sept. 1971, S. 178-184.

[3] *H. Heuser*: Unendlichkeiten, B. G. Teubner Verlag, Wiesbaden 2008, S. 30.

Friedrich Herrmann

4.16 Entropy as a measure of irreversibility

Subject:

Sometimes entropy is introduced as a measure of the irreversibility of a process. In this way it is possible to get a certain intuitive idea of the entropy which otherwise has the reputation of being an abstract and obscure quantity.

Deficiencies:

1. When introducing entropy as a measure of irreversibility, the question gets easily out of sight what happens with the entropy after it has been produced. A related question is what is the effect of that entropy whose origin we do not know. The entropy of the universe is constant in very good approximation. The entropy production which we observe in our immediate surroundings seems important to us but it is insignificant on a cosmic scale. Even on a terrestrial scale the produced entropy plays only a minor role in the total entropy balance. The entropy contained within the terrestrial globe is a about a million times that which is produced in a year at the Earth's surface (essentially by the absorption of the sunlight). For whom knows entropy only as a measure of the irreversibility of a process this entropy does not mean much.

2. We encounter entropy in the equation

$$P = T \cdot I_S \,. \qquad (1)$$

The equation tells us that every entropy current is accompanied by an energy current. It has the same structure as the familiar relation

$$P = U \cdot I \,.$$

Equation (1) is useful for the description of heat engines. Their working principle is easy to understand: Entropy is flowing through the engine; the entropy current at the inlet is equal to that at the outlet. Inside of the engine the entropy goes from high to low temperature thereby doing work, i.e. the engine emits energy by means of the output shaft. For the description of the working principle of the engine we need the entropy. Since the process is reversible the interpretation of entropy as a measure of irreversibility is of little help.

3. When looking for a measure of irreversibility, entropy is not really a good choice. Imagine we want to compare the irreversibility of processes going on in systems A and B. What we want to compare is not states but processes. Therefore, a statement about the entropy of A and B is not useful. It is better to consider the entropy production rate of A and B. However, if the production rate is greater in A than in B, we cannot conclude that process A is "more irreversible". If system A is much larger than B, it can be that the process of B is "more irreversible". Thus, in order to get a more convenient measure of irreversibility we should relate the entropy production rate to the size of the system. A better measures would be the molar entropy production rate.

Origin:

The lack of an intuitive idea of the entropy that is stored in a system and the intention to add to the statistical interpretation a phenomenological interpretation.

Disposal:

When entropy is interpreted as the heat content ("heat" in the colloquial sense of the word) the molar entropy production rate is a trivial byproduct.

Friedrich Herrmann

4.17 Negative entropy

Subject:

In textbooks on biology and related areas one may find statements about negative values of the entropy, for instance: “Living systems steadily produce positive entropy. In order to escape a decomposition into the thermodynamical equilibrium, these systems need a continuous supply of negative entropy. The only abundant source of negative entropy, which is available to living systems, is the excitation energy of the pigments. The excitation is carried out by photons. The only natural source for photons is the sun.”

Sometimes negative entropy is also called negentropy, and negentropy, so it is said, is identical with Shannon’s information.

Deficiencies:

Such statements partly only offend physical practice, but partly they are incorrect.

1. Whenever the value of an extensive quantity X increases in system A and decreases in system B, because there is a current of X between A and B, there are two possibilities to bring the process into words: Either one says that there is a current of positive X from B to A, or one says that there is a current of negative X from A to B. The theory (more exactly: the continuity equation) does not distinguish between these two ways of speaking. Only in the case that a velocity can be unambiguously attributed to the current, i.e. the current density j_X can be expressed by a density ρ_X and a velocity v

$$j_X = \rho_X \cdot v \,,$$

such a distinction may be justified. If the density ρ_X is negative, it can be argued that it is convenient to say that negative X is flowing from A to B. If ρ_X is positive, one would prefer to say that positive X flows from B to A. Even so, this distinction is not necessary.

However, if the quantity X can admit only positive values in principle, as it is the case for mass or entropy, then it is certainly not convenient to say that negative X is flowing from A to B, since this kind of wording suggests that a negative density of mass or entropy exists.

When speaking about negative entropy, as in our citation, one clearly pursues an objective: One intends to attribute the merit for the fact that the entropy of the living system does not increase to the sun. And now comes the mistake: Even if one defies that there is no negative entropy, one should admit that the flow of the hypothetic negative entropy takes the same path as the positive entropy that is actually flowing, although in the opposite direction. Now, since the positive entropy is flowing into the environment, one might at best say, that negative entropy flows from the environment into the living system. Thus the claim that negative entropy comes from the sun, is simply not correct. Since the subject is rather complex, the fact has gone unnoticed.

To further illustrate the problem let us apply the statement to a system which is more transparent: the heating wire of an electric heater. The normal and correct description of the entropy balance would be as follows: In the wire entropy is produced. This flows out into the environment. A statement that is analogue to the one of our citation would say: With the electric energy negative entropy is supplied to the wire. This entropy is compensated by that entropy which is produced in the wire. Obviously this statement is not correct.

2. If negative entropy (or negentropy) is identified with information (amount of data) then one is making a mistake of another kind. Let us first remind that entropy S and information H are calculated by means of the same statistical formula:

$$S = -k\sum_i p_i \ln p_i \qquad\qquad H = -f\sum_i p_i \ln p_i$$

Here p_i is the probability of the system to be in the microstate with the number i. k is the Boltzmann constant and f is a constant factor that is chosen in such a way that the unit of H is the bit. Thus, the values of both quantities are determined by the same procedure, what means that the two quantities are identical. The entropy of a system and the information stored in its microstate are, apart from a constant factor the same physical quantity.

Now, often the following awkwardness can be observed. We consider a system A. Instead of saying that the information H is stored in A it is said that H is the information that the observer is missing. And one goes yet one step further: One says that the observer has negentropy N with:

$$N = -H \,.$$

Instead of attributing the value to the system for which it was calculated or measured, one takes the negative of this value and attributes it to the complement of the system, i.e. to the environment or to the observer who is a part of the environment. It is as if you described the mass m of a body saying that the environment has the “negmass” $n = -m$. Such a procedure can certainly be kept up for a while, but there is no doubt that it is extremely uncomfortable.

Origin:

Negative entropy has a long tradition. Peter Guthrie Tait, a thermodanymicist and friend of Lord Kelvin, already has thought of introducing a negative entropy but prescinded from it [1]: “It is very desirable to have a word to express the Availability for work of the heat in a given magazine; a term for that possession, the waste of which is called Dissipation. Unfortunately the excellent word Entropy, which Clausius has introduced in this connection, is applied by him to the negative of the idea we most naturally wish to express.”

Negative entropy was definitely introduced into physics by Schrödinger. In his book “What is life?” from 1944, which is devoid of any mathematics, he writes: “What is this precious something in our food which saves us from death? This is easy to answer. Every process, every event, everything that happens – you can call it as you want – in short, everything that goes on in nature, means an increase of the entropy of that part of the world in which the process takes place. Thus a living organism continually increases its entropy – or if one prefers, it produces positive entropy – and thus strives for the dangerous state of maximum entropy, which means death. It can only keep away, i.e. it can live only, by constantly extracting negative entropy from its environment, – which actually is something very positive, as we shall see soon. What is nourishing an organism is negative entropy. Or, to put it somewhat less paradoxical, it is the essence of the metabolism that the organism succeeds in getting rid of the entropy which it has to produce as long as it lives.”

These statements of Schrödinger provoked the objection of his colleagues. Schrödinger defended himself, but somewhat half-heartedly.

The term negentropy had been introduced in 1956 by Brillouin [2]. At that time there appeared several publications about the relation between the thermodynamical quantity entropy and the quantity *information*, introduced shortly before by Shannon. Brillouin attached so much importance to this idea that he called it the “negentropy principle of information”. As mentioned earlier, the inconvenience of his proposal is that he attributes the quantity information to the observer or experimenter and not to the system for which it is calculated.

One reason of this inconvenient assignment may be the name that is usually given to the quantity: information.

Suppose for a computer data store (of for the microstates of a perfect gas) it has been calculated: $H = x$ MByte. When calling the quantity H information it seems logical to say: “I lack the information x MByte about the data store (or about the perfect gas).” If on the contrary, H would be called *amount of data*, then another wording seems more appropriate: “The amount of data within the data store or in the perfect gas is x MByte.” Thus when using the term amount of data the values of the quantity are correctly attributed to the data store or to the gas.

Disposal:

1. There is no need for a negative entropy. Everything is clearer when one is content with the positive entropy.

2. Attribute the quantity H to the data store (or the thermodynamic microstates) and not to the observer. Call it *amount of data*.

[1] *P. G. Tait*: Sketch of Thermodynamics, Edmonston & Douglas, Edinburgh 1868, p. 100.

[2] *L. Brillouin*: Science and Information Theory, Academic Press, New York 1962, p. 152f.

Friedrich Herrmann

4.18 Entropy and life

Subject:

Biological systems are highly ordered systems that form spontaneously. This fact is often considered a problem. One might believe, so it is said, that the genesis of a living organism is in contradiction to the second law of thermodynamics, according to which the entropy in a closed system cannot decrease. In reality, we are told, the second law is not violated, since biological systems are open systems. Their entropy can indeed decrease if thereby the entropy of the environment increases.

Deficiencies:

First we are told that there is the problem that the order of a biological system increases spontaneously. Sometimes this is even presented as a kind of paradox. And then the mystery is unravelled: We can calm down, all is right.

However, one could have put an end to the subject at an earlier stage, before it became a problem or a paradox. Actually the entropy content of biological systems is in no way uncommon.

A human being for instance consists of about 60 % water. At a temperature of 25 °C the entropy content of water is 3.9 J/(K · g). At the temperature of the living human body it is a little more. The remainder of the body is formed by proteins, lipids and carbohydrates whose entropy per mass is not very different from that of other condensed organic substances. It is thus 1 to 3 J/(K · g). It is not difficult to imagine a non-biological system which coincides with the human body not only in mass, volume and temperature, but also in entropy. Thus, the entropy content is not characteristic for a biological system.

When comparing the human being with a pile of sand of the same mass and the same temperature, it performs even more poorly, as far as entropy is concerned. Since the sand is a crystalline material its entropy is only about a forth of that of the person.

By the way, when a biological system grows its entropy does not decrease but increase. This is simply because its mass increases. If a person's weight increases by 2 kg, her entropy increases by about 4 kJ/K.

In conclusion, entropy is no more a characteristic of a living system than mass or volume.

Origin:

Probably various factors add up. First: The nonscientific connotations of entropy. One seems to believe that entropy has to behave characteristically in the context of such complex processes as organic life. Second: The lack of knowledge of the values of entropy (which by the way are easily found in tables). Third: The introduction of the entropy via statistical physics. The method is fascinating but apparently inappropriate for getting an idea about the entropy's values.

Disposal:

Introduce entropy as a measure of the quantity of heat. Do it in such a way that the student gets an idea of its values. Then the assumption that the entropy of a living system might display any peculiarity doesn't come up from the beginning.

Friedrich Herrmann

4.19 The Carnot cycle

Subject:

In the physics lecture at the university students get to know the Carnot cycle. An ideal gas goes through a cyclic process in four steps: two isothermal and two isentropic (or adiabatic) sub-processes. The students learn that in such a process only a certain part of the heat can be transformed into mechanical work. Sometimes the considerations are generalized to an arbitrary cyclic process by decomposing it into infinitesimal isothermal and isentropic parts. Usually, the process is represented in a *p*-*V* diagram.

Deficiencies:

1. Carnot's work consists in the presentation and substantiation of a single great idea: in a perfect heat engine the “caloric” gets from a higher to a lower temperature and thereby does work (in the french original “puissance motrice”). The work is proportional to the temperature difference and the amount of caloric. According to Carnot the working principle of the heat engine is analogue to that of a water wheel, in which the water passes from a higher to a lower level while doing work. Carnot introduces this idea at the beginning of his work. Only then he explains how we can imagine the details of a thermal engine. All what the students learn today about Carnot’s ideas is the tedious calculation of a special realization of a heat engine.

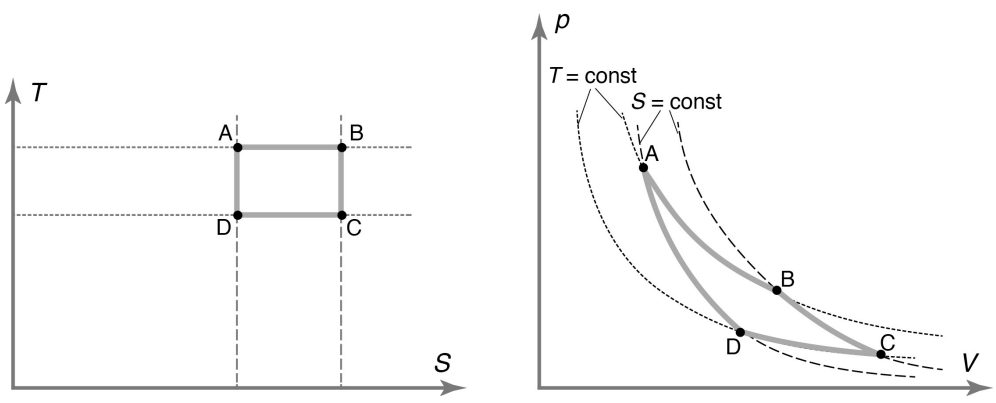


Fig. 1. *T*-*S* and *p*-*V* diagram of the Carnot cycle. In one cycle the engine receives entropy at a high temperature and gives it away at low temperature.

2. Two energy forms are involved in the Carnot cycle: *pdV* and *TdS*. If one really wants to treat the Carnot cycle it is best to represent the process in the coordinates of both energy forms: in a *T*-*S* diagram and in a *p*-*V* diagram, figure 1. The *T*-*S* diagram shows how simple the process is. Here Carnot’s idea can be seen clearly: the caloric (today we call it entropy) enters the engine at a constant high temperature (process step AB) and leaves it at a constant low temperature (process step CD). The *T*-*S* diagram is the same, whatever is used as the working substance. Not so the *p*-*V* diagram. This is an important fact which Carnot emphasizes before he discusses the special case of the “ideal gas”.

But we can also, as Carnot himself, abstain from both diagrams. Indeed, the *T*-*S* diagram is trivial, whereas the pV diagram is not interesting.

3. It is easier to get an understanding of heat engines by considering a continuous-flow machine instead of a machine that employs a cyclic process.

Origin:

1. Since in the fifties of the 19th century, in the context of the euphoria that accompanied the introduction of the energy, caloric was interpreted as an energy form Carnot's ideas appeared to be partially wrong. His caloric runs through the thermal engine. On its way from the high to the low temperature its amount does not change. From a modern point of view this behavior complies with that of the entropy. We can thus identify Carnot’s caloric with the entropy which was officially introduced in 1865 by Clausius. When interpreting caloric as energy, an incorrect conclusion results: At the low-temperature exit of the machine less energy comes out than what enters at the high temperature inlet. (The difference corresponds to the work done by the machine.)

The unsuitable interpretation of Carnot's idea has survived to this day, although in the course of the further development the error had repeatedly been pointed out.

2. A related concern is that entropy has survived as an unintuitive quantity. Thus, teachers try to avoid to deal with it whenever this is possible. Thereby the simple *T*-*S* diagram gets left behind.

3. At Carnot’s time cyclic processes were the norm, since the only thermal engine known at that time, the steam engine, was a cyclic machine. Steam turbines did not yet exist. The water wheel, to which Carnot refers was a continuously working machine.

Disposal:

What has this subject to do with school physics? Isn’t it clearly a topic for the university? Indeed, when presented in the usual, complicated way, it goes beyond the scope of school physics. But if it is presented in the clear manner as it had been done by Carnot, then it suits perfectly for school.

Friedrich Herrmann

4.20 Carnot efficiency

Subject:

The relatively low efficiency of thermal engines is caused by the fact that heat can only partially be transformed into work. The fraction of the total amount of heat which can be transformed into work under ideal conditions is called the *Carnot efficiency*. If T_2 and T_1 are the temperatures of the reservoirs of the incoming and of the outgoing heat, respectively, the Carnot efficiency is $\eta = (T_2 - T_1)/ T_2$.

Deficiencies:

What appears as a peculiarity of thermal engines is nothing more than the expression of a strange inconsistency. For comparison, let us consider a mechanical example, which had already been used by *Carnot*: A mill, perhaps in the Black Forest, with a water wheel of a height $h_2 - h_1 = 5$ m and situated at $h_2 = 1000$ m above sea level, gets with every kg of water a potential energy $m \cdot g \cdot h = 10$ kJ. From this amount it can use, in ideal conditions, only $m \cdot g \cdot (h_2 - h_1) = 50$ J. Thus, the *"Carnot"* efficiency is

$\eta = (h_2 - h_1)/ h_2 = 0{,}005$.

An identical mill on the lower Rhine, say at $h = 20$ m above sea level, would have an efficiency of 0,25. Here, we have generously calculated the potential energy against sea level. Compared to the approximately 20 MJ of potential energy when calculated with respect to the center of the earth, the resulting efficiency for both mills gets really depressing: $\eta = 2{,}5 \cdot 10^{-6}$.

One feels immediately that something is wrong here. Apparently, the Carnot efficiency has nothing to do with either the mill, or with the steam engine, but only with the position of the effective levels h_2 and h_1 or T_1 and T_2, with respect to the fictitious reference level.

Origin:

S. Carnot, who wrote down his ideas before the energy principle was formulated, did not know the quantity . He compared the steam engine with a water mill. For him the work originated, just as in the case of the water mill, from the difference of the "potential energy" of the heat in the reservoir of the incoming and the outgoing heat.

Disposal:

In thermodynamics, the concept is as superfluous as in mechanics. Described as an "entropy mill", the heat engine is just as trivial as a watermill.

Georg Job

4.21 Efficiency and Carnot factor

Subject:

The efficiency of a machine is defined by the quotient of the delivered useful energy to the total supplied energy:

η = useful energy/supplied energy

For a normal resistance heater, we obtain with this formula an efficiency of $\eta = 1$.

With a thermal engine one puts for the denominator of this expression all the energy which comes from the heat source and flows into the actual engine. If the thermal engine is ideal, i.e. if no entropy production takes place, the efficiency turns out to be the so-called Carnot factor:

$\eta = (T_2 - T_1)/ T_2$

With a heat pump, one puts for the energy delivered in the desired form the energy which leaves the heat pump at the high temperature T_2, and obtains:

$\eta = T_2 /(T_2 - T_1)$

Deficiencies:

The efficiency is awkwardly defined. One expects from a reasonably defined efficiency that

1. its value lies between 0 and 1;
2. an ideal machine has an efficiency of 1;
3. a non-ideal machine has an efficiency < 1.

A machine is ideal when it works reversibly, or in other words: if no entropy is produced.

None of these three criteria is fulfilled by the definition of the efficiency indicated above. The efficiency of the heat pump is greater than 1, and the first condition is not met. The ideal, thus reversibly working Carnot machine has an efficiency which is less than 1, so the second condition is not met. The resistance heater, which works non-reversibly and is a notorious energy waster, has an efficiency of 1. Thus the third condition is not met.

Origin:

The search for the definition of an efficiency, an effectiveness or an economic coefficient of thermal machines accompanied for nearly a hundred years the intricate process of differentiating between energy and entropy. It is not found in Carnot's work. Carnot probably would not have appreciated the now common definition. We found it in the work of Helmholtz but we do not know for sure whether Helmholtz is the inventor of this measure.

Although it was an unfortunate choice from the beginning, it is at least understandable why such a definition was made. On the one hand heat pumps did not yet exist, i.e. machines that have, according to the above definition, an efficiency grater than 1. On the other hand, there were still no fuel cells, and it seemed that the only way to take profit of the energy of carbon was to burn it. Therefore, it did not matter whether one attributed the low efficiency of a steam engine to the furnace or to the actual machine.

Disposal:

One uses the following definition for the efficiency:

$\eta = P_{ideal} / P_{real}$

P_{real} is the energy consumption of the real machine, whose efficiency one would like to evaluate. P_{ideal} is the energy consumption of a machine or a plant that performs the same task, but works in a reversible way, i.e. without entropy production.

With this definition, one obtains $\eta = 1$ for the reversibly working Carnot machine, because it is identical with the ideal machine of the same performance.

For the heat pump one always obtains a value of η which is smaller or equal to 1. If the machine works without any losses, that is, without friction losses, heat losses or losses in electric conductors, then it is ideal, and the efficiency will be equal to 1. In the case where we have such losses, will be less than 1.

A resistance heater supplies a certain entropy current (heat current) I_s at the high temperature T_2. The corresponding ideal machine is a heat pump, which supplies the same entropy current I_s at the same temperature T_2. It receives this entropy from the environment at temperature T_1. Thus, its energy consumption is

$P_{ideal} = (T_2 - T_1)\cdot I_S$

On the other hand the energy consumption of the resistance heater, which gives the same heat current $T_2 \cdot I_s$, is

$P_{real} = T_2 \cdot I_S$.

For the efficiency we obtain

$\eta = P_{ideal}/P_{real} = (T_2 - T_1)/ T_2$

i.e. is equal to the Carnot factor.

The resistance heater is wasting the more energy the higher the ambient temperature T_1 is. Indeed, the higher T_1, the lower the energy expenditure for raising the entropy from ambient temperature to the desired value by means of the heat pump.

On the basis of the same consideration, for any other irreversible heater we obtain for the efficiency the Carnot factor, for example the furnace of a coal-fired power plant. Thus, the "weak point" in such a plant is not the nearly reversible turbine, but the irreversible working of the furnace.

The definition given here is known in thermodynamics as "second law efficiency". It is introduced as an advanced concept. We propose to use from the beginning this definition, and call it simply "efficiency".

Friedrich Herrmann

4.22 The zeroth law of thermodynamics

Subject:

"If two systems are each in thermal equilibrium with a third, they are also in thermal equilibrium with each other." This statement is called the Zeroth law of thermodynamics.

Deficiencies:

If two systems are in thermal equilibrium, their temperatures are equal, and if their temperatures are equal they are in thermal equilibrium. From this fact follows the Zeroth law. There is no doubt that the statement is correct. However, it represents such a simple conclusion that it is hard to understand, how it could reach the state of a "Law of thermodynamics".

Who reckons that there is a profound meaning hidden behind the words, should remember that several other statements about other equilibria could be formulated which nobody would call a "Law" of anything, since the content of this statements is obvious.

Since the Zeroth law is often cited in the context of statistical thermodynamics, we shall, for comparison, consider the chemical equilibrium. In statistical thermodynamics, the chemical potential plays a role that is rather similar to that of the temperature: Together with temperature it is one of the two parameters in the probability distribution of the energy. Thus, in addition to the Zeroth law we could formulate an analogue "law" for chemical equilibria:

"If two systems are each in chemical equilibrium with a third, they are also in chemical equilibrium with each other."

Phenomenological thermodynamics shows us that we can formulate yet various other "Zeroth laws": One for each of the terms in Gibb's fundamental equation:

$$dE = TdS - pdV + \mu\, dn + \boldsymbol{v}d\boldsymbol{p} + \boldsymbol{F}d\boldsymbol{s} + \boldsymbol{\omega}d\boldsymbol{L} + \psi dm + \varphi dQ + Id\phi \ldots$$

(T = absolute temperature, S = entropy, p = pressure, V = volume, μ = chemical potential, n = amount of substance, $\boldsymbol{v}$ = velocity, $\boldsymbol{p}$ = momentum, $\boldsymbol{F}$ = momentum flow, $\boldsymbol{s}$ = displacement, $\boldsymbol{\omega}$ = angular velocity, $\boldsymbol{L}$ = angular momentum, ψ = gravitational potential, m = mass, φ = electric potential, Q = electric change, I = electric current, ϕ = magnetic flux)

So we could formulate for three bodies which by means of inelastic collisions attain the same velocities:

"If two systems are each in velocity equilibrium with a third, they are also in velocity equilibrium with each other."

Origin:

The need for the formulation of the Zeroth law seems to arise, when temperature and chemical potential are introduced in statistical mechanics. Then it has to be shown that one of the parameters in the probability distribution has the property of that quantity which is familiar to us and which we call temperature. However, even in this context the Zeroth law is no more than the expression of the transitivity of a physical quantity.

Disposal:

We do not treat the Zeroth law at school. What then is the relevance of the subject for school physics? It helps us to understand why thermodynamics is so unpopular at university and at school. With no other intensive quantity we make such a great play as with the temperature, with no other extensive quantity we make such a big fuss as with entropy. Sometimes, thermodynamics may remind the emperor's new clothes.

Regarding the disposal of the Zeroth law in particular, we have to ask colleagues from the university. We recommend our students: Don't allow to persuade you that there is a problem where there is none.

Friedrich Herrmann

4.23 The Third Law

Subject:

"It is impossible to reach the absolute zero of temperature by any finite number of processes."

This is one of several possible formulations of the Third Law of Thermodynamics.

Deficiencies:

Why do we believe that this statement is worth mentioning? There is a great number of other statements of the impossibility of something. It is impossible to empty an air-filled recipient by a finite number of processes. It is impossible to scoop a bath tub completely by means of a bucket. We perceive statements of this kind as trivial. We do not number them among the laws of nature. It is different, however, with entropy. We get to know it only in such an esoteric "wrapping", that an unprejudiced handling of it is extremely difficult. Statements about the entropy become a significance that is not in relation to its simple physical properties. We pay so much deference to the entropy and attribute to it so many metaphysical connotations, that our comparison with the emptying of a bath tub may seem disrespectful. Yet both statements are of the same kind. Our simple analogy describes the situation in a clearer way as all of the current formulations of the Third Law.

Origin:

The statement goes back to *W. Nernst*. His scholar *F. Simon* formulated the Third Law in the following way: "It is impossible to completely deprive a substance of its entropy." The statement compensates a deficit let by the Second Law, since it determines the constant of integration when calculating the entropy.

Disposal:

Our respect for the creators of the Third Law should not prevent us from seeing things a little more soberly. The place of the statement should not be the altar, but the box of our standard tools.

Georg Job

4.24 Microscopic – macroscopic

Subject:

According to many people entropy is a difficult quantity. They argue that a true understanding is only possible when its microscopic meaning is understood: Entropy is a certain characteristic of a probability distribution or, somewhat more intuitively, a measure of the disorder in the occupation of the microstates of the considered system.

Deficiencies:

1. Classical or "phenomenological" thermodynamics is one theory, statistical thermodynamics another. A theory is a mathematical description of a certain class of natural phenomena, and usually there is more than one theory to describe the same phenomena. In general we cannot say that one of these theories is better than the other. One of them may be more convenient for one purpose, whereas the other is better for another purpose.

As an example consider theories about the light. One of these is geometrical optics, another is wave optics, a third one is the thermodynamical description of light and a forth one is quantum electrodynamics. Each of these theories has its *raison d'être*. Nobody would claim that we no longer need geometrical optics since we have quantum electrodynamics. Quantum electrodynamics would be of no avail for the calculation of a photographic lens.

In the same way phenomenological thermodynamics is not better and not worse than statistical thermodynamics. To solve certain problems phenomenological thermodynamics is more adequate than statistical thermodynamics and for other problems it is the contrary.

2. Nature can be described on different scales of magnitude and of complexity. One might expect that the description becomes simpler when going to smaller scales. One might hope, that when penetrating into the microscopic world one gets closer to the indivisible, structureless, elementary particles. Up to now, however, experience taught us, that each time that we advanced one step further into the microscopic world, the searched-for elementary constituents of matter took a step back. At the same time we observe that, when going in the other direction, i.e. to the larger scales, the world does not necessarily get more chaotic and confusing, as one might have suspected, but that out of the complexity new simple rules and laws emerge. We learn from this, that it is not true that a microscopic description of the world is more fundamental than a macroscopic view, or in our particular case: phenomenological thermodynamics is not less fundamental than statistical thermodynamics.

3. For the description of thermal phenomena and its applications at a beginner's level phenomenological thermodynamics has an advantage over a microscopic approach. Entropy, when introduced suitably, turns out to be a quantity that is particularly intuitive. Its handling is so easy that a child gets along with it. One quickly and effortless comes to a quantitative description of thermal phenomena: Heat content, heat transport, phase transitions, thermal engines and efficiency. The three laws of thermodynamics appear as self-evident.

4. Nobody would claim that we understand an electric circuit consisting of a battery and a resistor only if we know the microscopic interpretation of electric resistivity, i.e. if we know about the electron-phonon coupling. Nobody would begin mechanics with the microscopic interpretation of mass by means of the Higgs mechanism. Ignoring the Higgs field does not prevent us from successfully applying Newtonian mechanics.

Origin:

For many scientists at the end of the 19th century the program of science was to reduce all physical phenomena to mechanics. There were good arguments to believe that this program was reasonable. It seemed natural to look for the mechanics that underly thermal, electric, magnetic and optical phenomena. In particular, mechanics seemed to govern the physical world on the microscopic scale. Everything seemed to be explainable by the motion and mutual interaction of small elementary "particles". Maxwell considered his electrodynamics a mechanical theory of the ether. With the kinetic theory of gases and statistical physics thermal phenomena could be traced back to the mechanics of the molecules or other particles. Only at the beginning of the 20th century it became clear that the non-mechanical theories were the more robust ones and that the mechanical interpretation of the world contained a certain part of fiction.

Disposal:

When teaching thermodynamics introduce entropy from the very beginning, in a similar way as you introduce mass in mechanics: As a quantity that can be easily measured directly, and for which we have a simple and direct intuition. Like mass is a measure of inertia and of weight, entropy is a measure for what in colloquial terms would be called the heat content. The concept of mass that is introduced in this way at school is suitable at all levels of teaching up to the University. Correspondingly, entropy as introduced in the above-mentioned way is a sound bases of university thermodynamics and technical thermodynamics.

Friedrich Herrmann

4.25 Temperature and kinetic energy of particles

Subject:

"When supplying heat the kinetic energy of the gas particles increases. Temperature is a measure of the time average of the kinetic energy of a particle. […] By the relation between energy and absolute temperature the concept of temperature gets an intuitive interpretation."

Deficiencies:

1. Temperature is not an intuitive concept for somebody who does not know about the kinetic energy of the particles? The physical layman has definitely no problem with the concept. The layman has a feeling of hot and cold, and is accustomed to the fact that a quantitative measure is used to describe this being hot or cold. Only a physicist can believe that the idea of the moving particles can help our intuition – obviously a case of "déformation professionnelle". Let us also remind that the swarming and wiggling particles which are supposed to help the student to understand temperature are themselves only a rough model, since the various excitations of liquid and solid substances are only badly in line with the idea of a particle.

2. It is suggested that temperature and the kinetic energy of the particles are up to a constant factor the same quantity. But they are not. In statistical physics temperature is a parameter in Boltzmann's energy distribution function and it does not matter in which degree of freedom the energy is stored. Translational motion is only one of many. Energy is also stored in rotations and oscillations of the molecules, in electronic excitations, in the various states of ionization, in plasmonic and magnetic excitations etc. So one might argue that the translational motion can at least be used as an indicator of the temperature. That is true, but why should we prefer the translation? We do not see it any better than the other degrees of freedom. We see it indirectly by the Brownian motion, but we also see other excitations indirectly: electronic excitations when the body is glowing or vibrational excitations by its infrared emission.

Origin:

Probably an afterglow from the old dispute about the nature of heat that was under way at the end of the 18th century. The question was: Is heat a substance or is it only movement [1]? Since the substance theory was rejected the interpretation of heat as the movement of the particles of matter was left. In the middle of the 19th century it was interpreted as an energy form. Since 1911it was also identified with entropy [2].

Disposal:

We can say that the kinetic energy of the particles is proportional to the temperature. But we should make clear that this is not the only way in which temperature manifests itself on the microscopic scale, since when temperature increases all the "energy stores" are getting more filled. Regarding our intuitive understanding of temperature: There is no need to help with a complicated microscopic interpretation.

[1] *J. Black, M.D*: Lectures on the elements of chemistry, edited John Robinson, LLD., Vol I, Edinburgh: Mundell and Son (1803), S. 30 -34

[2] *H. L. Callendar*: Proc. Phys. Soc. London **23** (1911), S. 153

Friedrich Herrmann

4.26 Entropy of mixing

Subject:

Two different gaseous substances are in two containers. When connecting the containers the gases intermingle. In this process the total entropy of the compound system increases. This increase is called entropy of mixing.

Deficiencies:

The choice of the name “entropy of mixing” is inappropriate and superfluous. The term suggests that the mixture of the gases has more entropy than the gases taken separately. Such a claim would presuppose that two systems are compared with one another: the mixed gases with the unmixed gases. Such a comparison could be done in three different ways according to how one imagines the mixing to be realized.

1. We start with two gases A and B in two containers, Fig. 1a. The container with volume V_A contains the amount of substance (measured in mol) n_A, the container with volume V_B contains the amount n_B. We connect the containers in such a way that we get a container with the volume $V = V_A + V_B$, Fig. 1b. The gases intermingle and the entropy increases by:

$$\Delta S = n_A \cdot R \cdot \ln\frac{V}{V_A} + n_B \cdot R \cdot \ln\frac{V}{V_B} \qquad (1)$$

This expression represents what is usually called the entropy of mixing. However, this increase of the entropy has nothing to do with the process of mixing of the gases. It is nothing else than an isothermal expansion of each of the two gases from its initial volume V_A and V_B, respectively, to the same final volume V. The entropy increase for such an expansion is

$$\Delta S_A = n_A \cdot R \cdot \ln\frac{V}{V_A}$$

for gas A and

$$\Delta S_B = n_B \cdot R \cdot \ln\frac{V}{V_B}$$

for gas B. Thus, the total entropy increase due to this expansion is equal to ΔS in equation (1).

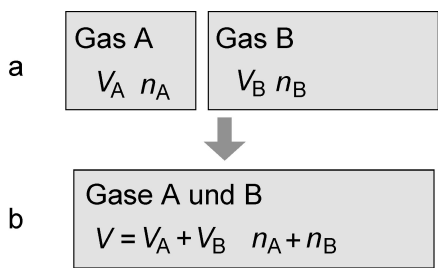


Abb. 1. Gas A expands from volume V_A to V, gas B from V_B to V.

2. We try another interpretation of the term “mixing”. We start with two containers of the same volume V, Fig. 2a, containing the two gases A and B, respectively. We compare with the situation of Fig. 2b, where both gases occupy a container of volume V. Here the total entropy before the “mixing” is equal to the entropy after it. Nothing is left to be called entropy of mixing.

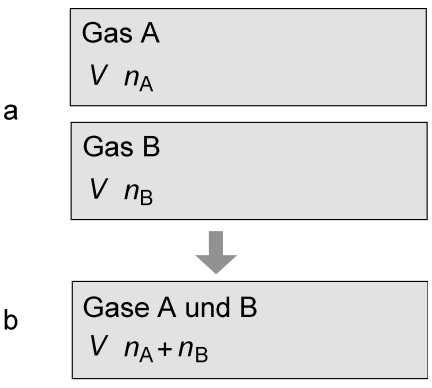


Abb. 2. Gas A and Gas B are transferred into the same container without changing their volumes.

3. Finally a third tentative, Fig. 3: We compare the entropy of gases A and B (amounts of substance n_A and n_B as before) both within the same container of volume V, with the entropy of a single gas C whose amount of substance is $n_C = n_A + n_B$. We ask for the difference of the entropies. What is the effect of taking away a characteristic that distinguishes between the gases A and B? Such a difference might also be considered a candidate for the name “entropy of mixing”. In this case, however, the entropy difference depends on the chemical nature of the gases and thus cannot be equal to the value given by equation (1).

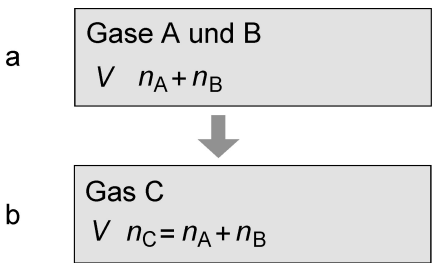


Abb. 3. Gases A and B are replaced with the same amount of a single gas C.

Origin:

Probably, when coining the term, one had in mind the idea that entropy can be considered a measure of the disorder of a physical system. The interpretation is correct, but its handling is not always easy.

Disposal:

Who knows that at constant temperature entropy increases with volume (and at constant volume with temperature) no longer needs the term.

Friedrich Herrmann and Peter Würfel

4.27 The Maxwell speed distribution

Subject:

The distribution function of molecular speeds in a gas calculated by Maxwell admits the value zero for $v = 0$. For increasing speed values it goes over a maximum value and tends again to zero for $v \to \infty$. The most probable speed v_{mp}, the mean value of the speed $\bar{v}$ and the root mean square value v_{rms} are different. The distribution is measured by means of a molecular beam. It can be visualized in a model experiment: Small moving spheres that exit a model gas, are classified according to their speed.

Deficiencies:

1. The aspect of the curve of Fig. 1, that is usually called Maxwell distribution might surprise. It is to be expected that high speeds are seldom, as the diagram shows. But shouldn't the probability increase the smaller the speed? In many texts this manifest question is not discussed. Actually, this behavior of the function can be considered an artifact of an inconvenient representation.

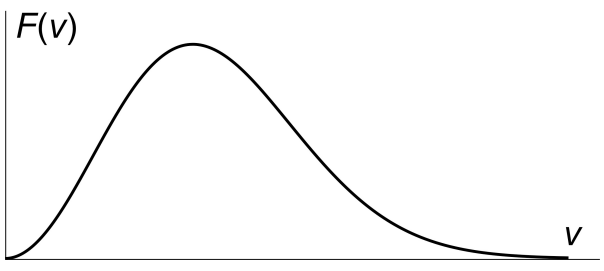


Fig. 1. Distribution of the absolute value of the velocity

The figure shows the distribution of the speed, i.e. the absolute value of the velocity. Velocity is a vector quantity. The laws of mechanics get complicated and clumsy when formulating them for the absolute values of the mechanical quantities (velocity, momentum, force). If in our case, we do not ask for the probability of finding a molecule with a speed in a given interval dv, but for finding a molecule with a velocity vector in the interval $dv_x\, dv_y\, dv_z$, we get a Gaussian distribution centered at $\boldsymbol{v} = \mathbf{0}$ (zero vector).

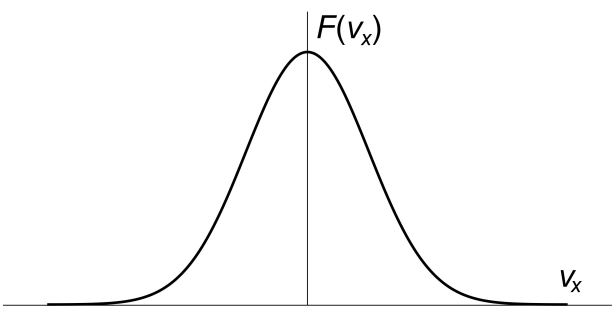


Fig. 2. Distribution of the *x* component of the velocity

Fig. 2 shows the probability distribution for one component of the velocity vector, see also equation (1) in Fig. 3. The reason why for $v \to 0$ the absolute values of the velocity (the speed) tends toward zero is that we do not compare equal volumes of the velocity space, but equal intervals dv. The volume $4\pi v^2\, dv$ in the velocity space that belongs to dv, increases for a given dv with the square of the speed, see equation (2) in Fig. 3. Hereby great speeds are "privileged" and small speeds are "penalized". In Maxwell's original work the two representations are clearly distinguished.

Particles with velocity vector $\vec{v}$

$$F_1(\vec{v}) = A \cdot e^{-\frac{mv^2}{2kT}} \qquad (1)$$

Particles with speed v

$$F_2(v) = B \cdot v^2 \cdot e^{-\frac{mv^2}{2kT}} \qquad (2)$$

Particle flow with velocity v in a molecular beam

$$F_3(v) = C \cdot v^3 \cdot e^{-\frac{mv^2}{2kT}} \qquad (3)$$

Particles with energy E

$$F_4(E) = D \cdot \sqrt{E} \cdot e^{-\frac{E}{kT}} \qquad (4)$$

Fig. 3. Various probability distributions. Constants have been merged into the factors *A, B, C* and *D* respectively.

2. It is often emphasized that the curve of Fig. 1 is asymmetrical, but usually it is not said what is meant by that: Is it the fact the curve itself does not have an axis of symmetry, or is it meant that it is not placed symmetrically to the axis of ordinates. It is also said that because of this asymmetry the values of v_{mp}, $\bar{v}$ and v_{rms} are different from one another. Sometimes it is insisted that one has to clearly distinguish between these values. The problem for the student is, that he or se does not know in which context this distinction is important. Most probably students will never get the opportunity to confound them.

3. Equations (1) and (2) make statements about the speed distribution of the molecules of a gas, that is in thermodynamical equilibrium. Often it is said or suggested that the same distribution hold for the particles in a particle beam, and that the distribution can be measured directly by means of such a beam. Actually, the speed distribution in a particle beam has a similar shape as that of the particle speed distribution in the case of equilibrium. However the functions are not the same, equation (3), Fig. 3. Here, the speed in front of the Boltzmann factor is at the power of three [1]. (For geometrical reasons there is a factor of v^2, but there is an additional factor v since a fast molecule contributes stronger to the current density than a slow one.)

4. Often, the importance of the speed distribution is emphasized without saying what it is needed for. The distribution of the components of the velocity allows to calculate the pressure. For many other purposes the distribution of the kinetic energy is needed. Also this distribution displays a certain similarity with the speed distribution, equation (4). It answers for instance the following questions [2]: "How many molecules of a gas have enough energy to initiate an endothermal chemical reaction, or to ionize an atom or molecule or to excite an atom, or escape from the gravitational field of the Earth or another planet, or to overcome the electrostatic repulsion between two atomic nuclei (in order to allow for a nuclear fusion reaction)?"

The only distribution that is not needed is that of the speed, i.e. the absolute value of the velocity.

Origin:

1. The speed distribution is found in Maxwell's work [3]. Maxwell's results are handed on from generation to generation, since Maxwell was a great physicist.

2. Maybe an effort to justify the claim that the average speed is a measure for the temperature.

3. The uncritical interpretation of the model experiment with the small spheres.

Disposal:

Other distributions are more useful, like that of the components of the momentum vector or of the kinetic energy. The model experiment would better be omitted.

[1] *W. Döring*: Einführung in die theoretische Physik V, Statistische Mechanik, Sammlung Göschen, Band 1017, p. 16.

[2] *H. Vogel*: Physik, Gerthsen - Kneser - Vogel, 13. Auflage, Springer-Verlag, Berlin, 1977, p. 169.

[3] *J. C. Maxwell*: On the dynamical theory of gases; Philosophical Transactions of the Royal Society of London, Vol. 157 (1867) p. 49-88

Friedrich Herrmann

4.28 Evaporating and boiling

Subject:

Liquid water (like other liquids) can transform into the gaseous state in two ways: It can vaporize and it can boil. If heat is transferred to water, first its temperature increases. When the boiling temperature is reached, its temperature does not go on rising. A corresponding differentiation does not exist for the process of melting.

Here, some typical comments taken from textbooks:

“When the vapor pressure gets equal to the pressure of another gas above the liquid, the liquid will boil. Then, the production of vapor not only takes place at the surface of the liquid, but also in the interior; vapor bubbles are forming.“ [1]

“A liquid boils, as soon as its vapor pressure equals the pressure of the air reposing on the liquid. The boiling temperature depends on the air pressure.” [2]

“Boiling: When boiling, in the interior of the liquid vapor bubbles are forming. As the liquid boils, its temperature does not change... Evaporating: The formation of gas takes permanently place at the liquid’s surface when the temperatures is below the boiling temperature.” [3]

Deficiencies:

There are no convincing answers to the following manifest questions:

1. Why does the process of phase change run slowly in the case of evaporation and fast in the case of boiling?
2. Why does the temperature not continue to increase, when the boiling temperature is reached?

It is easy to answer these questions:

The velocity of the process of evaporation depends on how rapidly the water vapor gets away from the water surface to places where the partial pressure of the vapor is smaller. This is a diffusional process and such processes are notoriously slow. Everybody knows that the process can be accelerated by blowing, i.e. by initiating convection. When the water is boiling the velocity of the evaporation is no more limited by diffusion. Since the vapor pressure at the surface of the water is now equal to the atmospheric pressure, i.e. the gas is pure water vapor, the vapor leaves the surface region not by diffusion but by a resistanceless streaming or flow process. The vapor can go away without any impediment. Now the evaporation rate depends only on the heating rate.

If entropy is delivered at a sufficiently high rate to liquid water with a temperature lower than the boiling temperature, the vapor that is produced cannot carry all this entropy away. The entropy accumulates and the temperature of the water increases. When the boiling temperature is reached this bottle neck disappears. The rate at which vapor is formed is now given by the rate at which entropy is delivered. This holds also true if the atmospheric pressure is not 1 bar, i.e. when the boiling temperature is less than 100 °C.

The formation of bubbles is eye-catching but it is not necessary for the process of boiling. If the water is heated from above with an IR lamp, boiling begins as soon as the upper surface reaches the boiling temperature, and no bubbles appear.

Origin:

The eye-catching phenomenon, i.e. the bubbles, seems to obstruct the view on the essential.

Disposal:

There is no way around considering the partial pressure of the water vapor above the water surface. As long as it is lower than the atmospheric pressure we have evaporation. The water escapes by the slow process of diffusion. When the water is boiling, the vapor above the water surface is pure water vapor. No resistance is opposing its evacuation. One explains the formation of bubbles but not without saying that this a an indication for the boiling process only when the water is heated from below. The explanation is yet easier when using the chemical potential, since both processes – the phase transition and the diffusion process – are driven by a difference of the chemical potential.

[1] Gerthsen-Kneser-Vogel, Physik, Springer-Verlag, Berlin, 1977, S. 189.

[2] Sexl, Raab, Streeruwitz, Das mechanische Universum, Band I, Verlag Moritz Diesterweg, Frankfurt, 1980, S. 205.

[3] Physik, GROSS-BERHAG, Ernst Klett Schulbuchverlag, Stuttgart, 1996, S. 92.

Friedrich Herrmann

4.29 Maritime climate and the heat capacity of water

Subject:

The conviction that the abnormally high specific heat capacity c of the water is responsible for mild winters and cool summers in countries near to a coast is part of the knowledge of a person with a general education in physics.

Deficiencies:

Why do we compare the heat capacities of 1 kg of the substances and not 1 mol or 1 m^3? In the present case neither the heat capacity per mass $C/m = c$, nor that per volume $C/V = c \cdot \rho$ is the adequate quantity, but at best the heat capacity per surface area $C/A = c \cdot \rho \cdot h$, where A is the surface and h the depth of the layer that is effective for the heat exchange. h is that depth up to which the seasonal temperature variations can be observed.

In order to estimate the value of h, we consider the heat Q, that is absorbed through the surface area A during the summer half-year $a/2$, where we admit a constant thermal conductivity λ and a constant temperature gradient $\Delta T/h$ within the layer: $Q \approx (1/2) \cdot a \cdot A \cdot \lambda \cdot \Delta T/h$. On the other hand, Q can be estimated from the heat capacity $C = c \cdot \rho \cdot h \cdot A$ and the average temperature difference $(1/2)\Delta T$ between the considered layer and the environment: $Q \approx c \cdot \rho \cdot h \cdot A \cdot (1/2)\Delta T$. Equalizing and separating h we get

$$h \approx \sqrt{\frac{\lambda a}{c\rho}}$$

and finally with $C/A = c \cdot \rho \cdot h$

$$\frac{C}{A} \approx \sqrt{\lambda c \rho J}$$

The table below shows that the water comes off badly in this comparison. The climate effectiveness of the water can only be understood when taking into account that in the ocean the roll-around of the water engages layers that are much thicker than the calculated 2m, and that the 1000 mm annual precipitation (typical for Europe) corresponds to a heat turnover that is sufficient to heat a 60 m layer of water by 10 K. What is decisive is that the water is liquid and that it is volatile, and not its abnormal specific heat capacity, which, due to the low mass density and the low thermal conductivity does not play an important part.

	ρ	λ	C/m	C/V	C/A	h
	Mg m^{-3}	J m^{-1} s^{-1} K^{-1}	kJ kg^{-1} K^{-1}	MJ m^{-3} K^{-1}	MJ m^{-2} K^{-1}	m
water	1,0	0,6	4,2	4,2	9	2
granite	2,8	3,6	0,8	2,3	16	7
basalt	2,9	2,1	0,9	2,5	13	5
river sand[1)]	1,6	1,1	1,0	1,7	7	5
ground soil[2)]	2,0	2,3	1,3	2,5	14	5

1) fine-grained with 0,07g/g humidity. 2) argillaceous soil with fine-grained sand and 0,14 g/g humidity

Origin:

We easily succumb to the temptation to conclude from a parallelism to a causal relationship and that even more as the conclusion is suggested by experts. The argument seems convincing and survives, because the refutation is not easy.

Disposal:

It is not enough just to elide such incorrect conclusions which are more frequent than is generally admitted.

Georg Job

4.30 The heat transport through the atmosphere

Subject:

For sunlight the atmosphere of the earth is almost completely transparent. Thus, the sunlight is not absorbed by the atmosphere but only by the surface of the earth. For the infrared radiation that is emitted by the earth the atmosphere is almost completely opaque. Seen from outer space, the IR radiation that comes from the earth is emitted by the atmosphere in a certain hight, called the emission altitude. The corresponding energy gets from the earth's surface to the emission altitude by various mechanisms, which are often represented in energy flow diagrams like that of figure 1. It is striking that there are two currents of radiant energy, whose magnitudes are larger than that of the energy flow that comes from the sun.

Deficiencies:

A representation like that of Fig. 1 suggests, that the most important contribution to the energy flow between the earth's surface and the upper troposphere comes from radiation. The corresponding arrows are the largest and the energy current density is the highest among all the other energy currents. This impression is supported by the text that accompanies such diagrams.

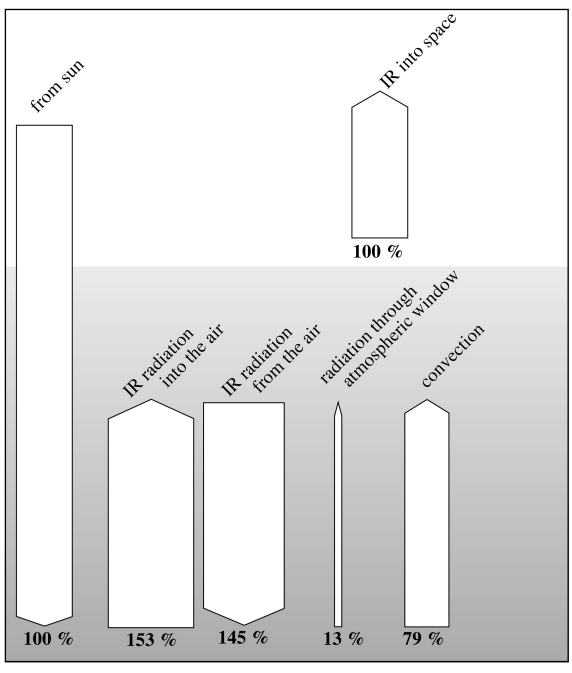


Fig. 1

The representation, though, is misleading. We obtain the net radiative energy flow between the earth's surface and the upper troposphere by adding the two radiative currents. The result is a current that corresponds to only 8 % of the energy current absorbed by the earth, Fig. 2. In this figure it is clearly seen that the dominant transport mechanism is not radiation but convection. For a rough treatment or a first approximation the radiation may

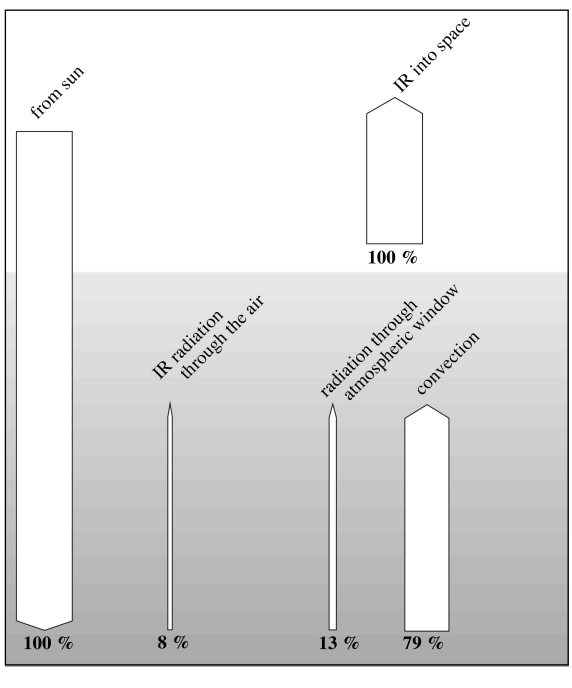


Fig. 2

even by completely ignored.

One might tend to defend the representation of Fig. 1 by arguing that it contains more information than Fig. 2. However, this information is in general not in demand, and as our experience shows, it disorients the students.

Actually, one can decompose any other current in two (or more) components. We can decompose the air at rest in the room where we are in two halves, the one consisting of the molecules that move to the right at a given instant of time, and the other half of those that move to the left. In this way we obtain two huge counterflowing air currents. In a similar manner one could proceed with the electrons within a currentless conductor: There would be two strong counterflowing electric currents. By the way, one could decompose in the same way the horizontal components of the IR radiation in the atmosphere.

Let us consider yet another example, that is even more similar to the radiative flow in the atmosphere: The heat flow within a solid body, say a copper bar that is heated at one end and cooled at the other. What the photons are for the IR radiation in the atmosphere are the phonons in the copper bar. It would not come to anybodies mind to decompose the phonon current in the copper bar into two counterflowing currents, when one is interested in the heat transport within the copper bar.

We do not claim that such a decomposition is incorrect. It is only complicated and misleading.

Origin:

Possibly the fact that it is easier to measure the component energy currents than the net energy flow.

Disposal:

Represent the undecomposed currents of the radiant energy in the atmosphere instead of the partial energy currents. This quantity has a well-defined value for each point of the radiation field. Then it will become clear that the dominant heat transport mechanism through the lower troposphere is convection.

Friedrich Herrmann

4.31 Shooting stars and space capsules

Subject:

The high temperatures that appear when a space shuttle, a space capsule or a meteoroid enters the earth's atmosphere are often attributed to friction.

Deficiencies:

Friction generates heat, and heat means high temperature. Everybody has made the corresponding experience. Brakes get hot, a blunt drill can be brought to incandescence, the blunt sawing blade of a circular saw causes the wood that is to be cut to carbonize. What is closer at hand than to explain the high temperatures at the entry of a space shuttle or a meteoroid into the atmosphere by friction? And it is still not correct.

A body that moves with a high velocity through the air, compresses the air at its front side. If the velocity is greater than that of the sound, in front of the body a shock wave is forming: a step change to higher values of pressure, density and temperature and lo lower values of velocity (in the reference system of the body). The higher the velocity of the body, the higher are these step changes. For a reentry capsule the temperature can attain 20 000 K. (Thereby the heat shield may heat up to 2000 K.) In the subsequent expansion and the accompanying friction and turbulence processes the temperature decreases again. Thus, the high temperatures appear in front of the "nose" of the flying object, and not at those places where the frictional processes occur.

Two questions arise:

– Why does the temperature increase within the shock wave?

– Why does the temperature not further increase in the subsequent frictional phase?

We obtain the answers to both questions by considering the entropy. All we have to know is, that the entropy of a given amount of a gas depends on its volume and on its temperature: the higher the temperature (for fixed volume) and the greater the volume (for fixed temperature), the more entropy contains the gas.

Regaring the space shuttle: We consider a certain portion of air. In the shock wave this amount of air is compressed very rapidly, and that means isentropically ("adiabatically"). Since the entropy does not change and the volume decreases, temperature goes up. The effect is the same as the temperature increase of a descending air mass in the atmosphere.

The second effect –the decrease of the temperature in spite of friction– may be more unexpected. In order to understand it consider a model system, that is geometrically simpler than the space capsule, the shuttle or the meteoroid, but in which the thermodynamical processes are essentially the same: a stationary flow of a gas within a tube with a flow resistance, Fig. 1. Since the gas expands when passing through the resistance the volume flow (the liters per second) behind is greater than that in front of it. If the gas can be considered as perfect, the process is described by a surprisingly simple equation:

$$c_p \cdot T + \frac{\hat{m}}{2} v^2 = \text{const}$$

Fig. 1

Here, c_p is the specific heat capacity at constant pressure, T the absolute temperature, $\hat{m}$ the molar mass and v the velocity of the gas. The equation tells us that the expression at the left side of the equation has the same value at every point on a streamline: When the velocity is high, temperature is low and vice versa*. (A condition for the validity of the equation is that the gas does not exchange energy through the walls of the pipe.) The equation remains valid when there is a flow resistance in the tube. For the case of Fig. 1 the equation tells us that the temperature behind the resistance is lower than in front of it. Since the gas expands, its velocity is higher behind the resistance.

When making the tube behind the resistance wider, it can be ensured that the velocities in front of and behind the resistance are equal, Fig. 2. Now it can be understood why the frictional process in the resistance does not result in a temperature increase. The entropy of the gas increases due to friction, but this increase does not manifest itself in a higher temperature but in a greater volume. Thus, the constancy of the temperature in the arrangement of Fig. 2 has the same explanation as that of the well-known Gay-Lussac expansion.

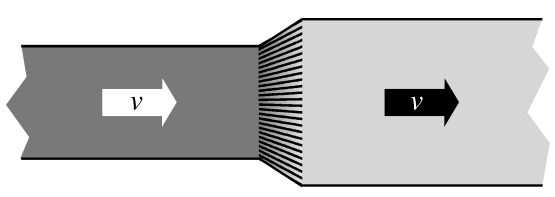


Fig. 2

The same arguments hold for the space shuttle. Here too we have an expansion together with a frictional process. In this case there is an additional effect: the air is mixing with the ambient air. The resulting effect is that the temperature of the air is actually decreasing**.

In summary: A gas when compressed isentropically gets warmer. When expanded its temperature decreases again. The temperature of a gas increases not only by friction but also by compression. The simple chain of arguments movement → friction → heat, which is correct for liquid and solid substances, is not correct for gases. Which property of a gas is responsible for this behavior? Simply the fact that gases are compressible.

Origin:

Temperature changes due to isentropic compression or expansion are ubiquitous effects. The most eye-catching manifestation is the snow on the high mountains. For many thermal effects the physical layman has an explanation which bear up against a thorough physical analysis. In the present case, however, such an explanation is missing: The laymen sees the snow at the top of the mountains, and he sees the correlation with the height, but not that with the pressure of the air. On the contrary he correctly explains temperature effects that are due to friction. So why not attribute the heat effect of the space shuttle or the meteoroids to friction? Apparently some physicists do not have an advance over the layman. The reason may be that they are not acquainted with the simply and powerful tool entropy.

Disposal:

Show that the isentropic compression and expansion are the cause of many striking thermal effects. The high temperature which causes the melting of the outer layers of a meteoroid and which endangers the space shuttle is only one of them.

* An example is the air coming out of a car tire when opening the valve. Within the tire the velocity of the air is zero, immediately after leaving the tire it is high. Therefore, its temperature decreases.
** Actually, the processes are even more complicated. Due to the high temperatures there are electronic excitations of the molecules, dissociations and other chemical reactions. All these cause a further decrease of the temperature.

Friedrich Herrmann

4.32 Thermal radiation

Subject:

We comment on three statements about thermal radiation, that can be found in textbooks. They are partly contradictory and partly incorrect. They appear not necessarily in the same textbook.

(1) Heat transfer takes place as heat conduction, convection and thermal radiation.

(2) At the red end of the spectrum of the visible light begins the domain of infrared or thermal radiation.

(3) With the light of the Sun heat comes to the Earth. It manifests itself by a temperature increase of the body that absorbs the light.

Deficiencies:

Our subject is that electromagnetic radiation, which is called *heat radiation* or *thermal radiation*. With these names one characterizes a particular method of creating electromagnetic radiation: A body emits electromagnetic radiation because its absolute temperature is greater than zero. There are other procedures for the production of electromagnetic radiation. The corresponding radiation is called non-thermal radiation. An example for non-thermal radiation are the microwaves produced by means of a klystron, the luminescence radiation from a semiconductor diode or the light from a laser.

First a somewhat trivial objection against one of the above-cited statements: Thermal radiation is not limited to the infrared domain. Sunlight is thermal radiation but most of its energy corresponds to the visible part of the spectrum. The cosmic background radiation is thermal, and its spectral maximum is situated in the microwave domain. The plasma of a fusion reactor emits thermal radiation in the X-ray domain.

A more serious and more subtle deficiency has to do with the statement that thermal radiation transfers heat. To analyze this statement we first have to clarify what is meant when saying that in a process heat or “energy in the form of heat” is transmitted. A heat transfer is an energy transfer that is accompanied by an entropy transport. The energy current P and the entropy current I_S are proportional to one another:

$$P = T \cdot I_S \, .$$

In general an energy flow has various contributions: heat, work, electric and chemical energy. Only that part is called heat, that corresponds to above equation.

To decide if or which part of a given electromagnetic radiation is heat we have to consider the energy flow and we have to know the temperature. If the radiation has a Planck spectrum, there is no problem. But the more selective the spectrum is, the more difficult it is to attribute a temperature to the radiation. The situation gets simple again when the radiation comes from a non-thermal radiator, as the microwaves from a Klystron or the electromagnetic waves from a radio emitter. In these cases that entropy flow is nearly zero and one will not call the emitted waves thermal radiation. When only considering the effect of the electromagnetic waves on an absorber, i.e. the fact that the absorber heats up, these difficulties are easily ignored.

We now come to an incorrectness in one of our citations. The heating effect of the absorber is not an indicator of the heat that may be transported by the radiation. The heating effect is due to the fact that the radiation transfers energy and that this energy is dissipated in the absorber. The “form” of the energy does not matter. A radiation that is completely entropy-free causes the same heating effect as thermal radiation, if both radiations carry the same energy and if both are completely absorbed.

Summing up: From the heating of an absorber we cannot conclude that the heat has been transported by the incident radiation.

Indeed, in those cases where the argument is most often used, i.e. in the case of the sunlight or the infrared radiation coming from a glowing body, the entropy production rate in the absorption process is much higher than the entropy inflow by the radiation.

We have seen that the heating of an absorber does not prove that the radiation transfers entropy. There is, however, a clear indicator for the entropy transferred by a given radiation. Instead of looking at the absorber, look at the emitter. If the temperature of a body, which has no material contact with its surroundings, decreases, i.e. if its entropy decreases, we can conclude that the radiation emitted by the body must have carried the entropy away, since the Second law tells us, that entropy cannot be annihilated.

Origin:

Heat radiation has been observed and studied long before it was possible to recognize that the nature of the radiation is the same as that of light, and long before the relationship between energy and heat was understood. The name radiating heat (strahlende Wärme) is probably due to Scheele [1]. In 1790 Pictet [2] believed that light and heat exist separately. In particular, he believed to have shown that moonlight is not accompanied by heat, whereas sunlight is. In the first decades of the 19th century the conviction was growing that light and heat radiation are of the same nature. However, for a definite clarification the appearance of two great theories had to be awaited: Maxwell’s electrodynamics and Planck’s statistical thermodynamics of radiation [3].

Disposal:

Do not identify infrared radiation and radiative heat transfer.

Regarding the electromagnetic radiation from the Sun, do say that it transfers energy. The heating of the absorber is mainly due to the dissipation of this energy. As an argument for the fact that radiation also transfers entropy consider the cooling of the emitter instead the heating of the absorber.

Instead of giving a name to the radiation – thermal radiation or heat radiation – attribute the name to the emitter: thermal emitter.

[1] *E. Mach*: Die Principien der Wärmelehre. – Verlag von Johann Ambrosius Barth, Leipzig 1919. – S. 126

[2] *M. A. Pictet*: Essai sur le feu. – Genève 1790

[3] *M. Planck*: Vorlesungen über die Theorie der Wärmestrahlung. – Verlag von Johann Ambrosius Barth, Leipzig 1913

Friedrich Herrmann and Peter Würfel

4.33 Sun and spectral lamps

Subject:

As a university student one gets to know two classes of light sources, that differ in their mode of operation. To the first class belong glowing bodies, the sun and the yellowish-white flame of candles. To the second belong spectral lamps, lasers, LED's and colored flames.

Regarding the light sources of the first class, the students learn that black, hot bodies emit electromagnetic radiation: the thermal or black-body radiation. The spectrum only depends on the temperature of the radiator. The corresponding function is called Planck's law.

Regarding the emission of the sources of the second class, the mechanism seems to be different: The students learn that electrons in atoms, molecules or in a crystal lattice go from an excited state to a state of less energy and thereby emit a photon. The frequency of the corresponding light is obtained from the energy difference between the two states; the intensity depends on the transition probability. The excitation can be realized in various ways: electrically, thermally or by optical pumping.

Deficiencies:

The various light sources are discussed on two different conceptual levels: the sun, the light bulb, the candle flame etc. are treated thermodynamically, the spectroscopic lamp, the laser and the LED are explained by discussing processes that go on at the atomic level. Thereby the impression may result that the emission of the light bulb has nothing to do with atomic physics, and that the emission of the sodium atoms when strewing table salt in a gas flame has nothing to do with thermodynamics.

In fact, both types of light sources are based on transitions of a system from an excited state to a state of lower energy, and in both cases the intensity is influenced by the laws of thermodynamics.

A justification for choosing two different approaches may be that one often is only interested in the shape of the spectra. Indeed, the spectrum of a black body can be obtained by only using arguments of statistical physics. The microscopic mechanism is irrelevant. On the contrary, thermodynamics does not help much in obtaining the spectrum of a spectral lamp. Here atomic physics is needed.

However, if one does not explain how the two patterns of explanation are related, one can only hope that the students don't do what normally we expect from them: ask questions when they do not understand. Such a question could be: Why does the sun not have a line spectrum like a hydrogen-helium spectral lamp?

Origin:

The theories and explanations of black-body radiation and of line spectra have been developed independently, and they have conserved this independence in the teaching of physics until today. Our example also shows that it is not at all clear what is meant by a "satisfactory explanation". In one case we present as an explanation the reduction to a microscopic mechanism, and in the other the description of the process of producing the radiation and the spectral properties of it.

Disposal:

A body whose temperature is not equal to 0 K emits electromagnetic radiation. If its emissivity $e(f)$ is equal to one at all frequencies (The values of e are between 0 and 1), the energy flow density j_E in the frequency interval df of the radiation is given by Planck's law:

$$dj_E = \frac{2\pi h}{c^2} \cdot \frac{f^3}{e^{\frac{hf}{kT}} - 1} df \quad .$$

Now, the emissivity of a body is for every frequency equal to its absorptance

$e(f) = a(f)$.

If e is equal to one for all frequencies, so is also a and the body is completely opaque; it is black.

If $e(f) = 1$ does not hold for every frequency, then Planck's law becomes:

$$dj_E = e(f) \cdot \frac{2\pi h}{c^2} \cdot \frac{f^3}{e^{\frac{hf}{kT}} - 1} df \qquad (1)$$

Since for each frequency e is smaller than or equal to one, the spectral energy flow density for a non-black body is for every f smaller than or equal to that given by Planck's formula. An example for such a spectrum is that of the light of colored flames: the bluish light of a methane flame, or the yellow light of a hydrogen flame in which some salt is strewed.

The fact that the spectra of glowing bodies can be described thermodynamically does not mean that the microscopic mechanism of emission is different in principle from that of a spectral lamp. Each photon that leaves an incandescent body is generated in a transition: for the photons of the visible light from an electronic transition, for the long-wave infrared photons from a vibrational or rotational transition. Since in a solid often transitions of all energies are possible, the special case of the Planck emitter is frequently realized. (However, not every macroscopic piece of matter must emit light with a Planck spectrum. A nice experiment that shows it is to heat two adjacent small pieces of different materials with a Bunsen burner; for instance a piece of iron on the one hand and a piece of quartz or sapphire or a white pebble on the other. Whereas the iron piece glows brightly, the quartz, sapphire or pebble does not emit any visible light at all.)

The sun represents a thermal radiator of a particular interest. We know that it consists almost exclusively of hydrogen and helium. Therefore one might expect that its spectrum is a line spectrum similar to that of a hydrogen-helium spectral lamp (which however would be excited thermally instead of electrically). On the other hand we know that sunlight has a continuous spectrum, which is quite close to a Planck spectrum.

How can these two statements be reconciled? The explanation comes from the factor $e(f) = a(f)$ in equation (1). For a hydrogen gas in the laboratory that is thermally excited this factor is almost exactly equal to zero for almost all frequencies. Except for some frequencies in the ultraviolet domain the gas is completely transparent. Therefore the emission spectrum of the lamp differs greatly from a Planck spectrum. However, the absorptance, and hence the emissivity of the body gets larger when the body gets thicker. When light enters a body and its path within the body is long enough, it will eventually find a suitable transition. The corresponding path length depends of course on the frequency of the light. In the sun it is at worst a few hundred kilometers. This is much when compared to the size of a spectral lamp, but it is very little when compared with the radius of the sun. A layer of solar matter with a thickness of 10 cm (taken from the area of the photosphere) is fully transparent: It practically does not absorb and thus not emit (in the direction perpendicular to the layer). As the thickness of the layer increases, the absorption, and therefore the emission, also increases. The spectrum is now similar to that of a spectral lamp [1]. When the thickness of the gas layer further increases, there is more and more emission in the region between the spectral lines. A layer with a thickness of 1000 km absorbs every light completely, and thus emits like a black body of a temperature of 6000 K.

But what are the transitions that are responsible for this absorption and emission? Even if we only had a pure hydrogen-helium mixture, and if we neglected the (weak) ionization, we would get total absorption in the visible domain when the layer has reached a thickness of several hundred kilometers. This absorption is due to the width of the spectral lines. But there are more absorption mechanisms: The hydrogen-helium gas is weakly ionized, and the free electrons absorb light. In addition, solar matter also contains all other elements, albeit in low concentration and the absorption of these elements partly lie in the visible spectral range.

Since we know that for our purposes the path length of the light within the sun can be considered as arbitrarily long we need not bother for the question of which of these is the dominant absorption mechanism. This is similar to the well-known experiment, consisting of a box into which a small hole has been cut. The hole is black, whether the interior walls reflect or scatter, whether they are black, white, yellow or blue. The absorption mechanism for the blackness of the hole is irrelevant.

[1] *M. Vollmer:* Hot gases: The transition from line spectra to thermal radiation. Am. J. Phys. **73**, (2005), p. 215.

Friedrich Herrmann

4.34 Temperature and heat of a gas when expanding into the vacuum

Subject:

In many respects the properties and the behavior of entropy coincide with those of what in colloquial terms is called heat. However, there is a famous experiment, where the two concepts seem not to match: the expansion of a gas into the vacuum, known as Joule-Gay-Lussac expansion.

Deficiencies:

Teachers who introduce entropy as a measure of the colloquial heat sometimes hear the objection that in the Joule-Gay-Lussac experiment entropy increases, but no „heat is created“ [1].

Various remarks can be made in this respect:

1. The correspondence or coincidence between a physical quantity and what its name seems to promise is never complete or perfect. An example is the quantity *Q* and its established name „heat“. For the student it is hard to accept, that it is not possible to say, that a body contains a certain amount of heat. The correspondence between the name and the physical meaning is even worse for the quantity *F*, called „force“.

2. In the present case it is not even the question for the correspondence between the name of a quantity and its meaning in physics, since the name of the quantity *S* is entropy. The question is only: Is it advisable to mention the concordance between the colloquial heat concept and the properties of the physical quantity entropy. Actually, this correspondence is better than it is for most of the other physical quantities.

Let us take as an example force. It is common practice to introduce the force by appealing to our „muscular sensation“. However, this „muscular sensation“ is no more characteristic for a force than for an energy current. Nevertheless, nobody takes offense at this comparison.

3. Back to the expansion into the vacuum. It is not an experiment that one would discuss at the beginning of the thermodynamics course, but only in the context of the treatment of gases. The discussion could run as follows: A gas is expanding into the vacuum. After thermodynamic equilibrium has established, the temperature is nearly the same as before the expansion. Do we have to conclude that the heat content („heat“ in the colloquial sense) did not increase? At first sight yes. However, it would be careless to judge the entropy content (or the content of the colloquial heat) before and after the expansion only by looking at the temperature, since the volume of the gas has changed. We therefore bring the gas back to its initial volume, and we do that in a way that we neither add entropy to the gas nor extract entropy from it. After doing so we notice, that the gas is warmer than before, its temperature is higher. It must contain more entropy (or heat in the colloquial sense).

Origin:

We suppose that the denegation of the idea that entropy reflects the properties of the colloquial heat has nothing to do with a possible lack of compliance between the two concepts. It rather seems that some people are desperately looking for examples, that show the limits of the model, motivated by the fear that they might discover that entropy is not as complicated a concept as they had believed all their lives.

Disposal:

Introduce entropy by associating it with the concept of heat in the colloquial meaning of the word. The match of the two concepts is better than that for most of the other physical quantities, that we introduce at school.

Friedrich Herrmann

[1] *M. Bartelmann, F. Bühler, S. Großmann, W. Herzog, J. Hüfner, R. Lehn, R. Löhken, K. Meier, D. Meschede, P. Reineker, M. Tolan, J. Wambach and W. Weber:* Expert opinion on the Karlsruhe Physics Course; commissioned by the German Physical Society

http://www.physikdidaktik.uni-karlsruhe.de/kpk/Fragen_Kritik/KPK-DPG%20controversy/Expert_opinion_english.pdf

4.35 Measuring entropy (add-on)

Subject:

Entropy has the reputation to be a difficult quantity. One of the reasons for this valuation seems to be the conviction that it is difficult to measure.

Deficiencies:

The knowledge of a measuring procedure is important for the understanding of a newly introduced physical quantity. The more transparent the measuring method, the better. One can also say: the more "direct" the measurement, the more concrete or vivid the conception of the quantity.

However, often the measurement method that is most easily understood is not at the same time the most exact or the most comfortable. Therefore, in order to get a clear understanding of a quantity it may be appropriate to introduce a measuring method that is not very precise and that is difficult to realize technically, but that is transparent and conceptually simple.

What about the entropy in this respect? Usually it is introduced in a way that is due to Clausius: "We attribute…to each state of a system a function *S* which we call the entropy of the state and whose complete differential *dS* for a reversible change is

$$dS = \frac{dQ}{T}$$

where *dQ* is the absorbed heat, *T* the temperature, at which the absorption takes place." [1]

From such a definition it can hardly be seen what kind of measurement has to be carried out in order determine the value of the entropy. How do we recognize if the change of the state of the system is reversible? How can we measure the amount of the absorbed heat? What do we have to do concretely?

Actually, measuring the entropy is very easy if one takes profit of its "producibility", i.e. the irreversibility [2]. The procedure is technically simple, inexpensive and precise. But is there a procedure that is conceptually more simple?

To answer this question we will ask how other quantities are measured which have an important property in common with the entropy: the property of being extensive or "substance-like".

For these quantities the measurement can in principle always be carried out in the following way: The amount of the quantity that is to be measured is transferred to the measuring instrument. The measuring instrument reacts with a corresponding deviation of a pointer. We are in this situation when measuring electric charge with an electrometer. The charge that is to be measured is transferred to the electrometer. The pointer of the electrometer shows a deviation that corresponds to the charge. The measurement is not precise, but an important property of the electric charge is clearly seen: Charged bodies exert forces on one another. And it gets obvious that charge is substance-like.

Momentum can be measured in a similar way: The momentum is transferred to the measuring device, which reacts with a kind of deviation or another visible signal. An example is the ballistic pendulum. Another procedure is described in [3].

Is it possible to measure an amount of entropy with a similar method? The corresponding device is shown in figure 1: A flask with a mixture of ice and liquid water. (A variant that is somewhat more complicated is Bunsen's ice calorimeter.) The entropy that is to be measured is transferred to the flask. As a consequence a part of the ice will melt. The amount of the melted ice is a measure of the entropy that has been supplied.

Fig. 1. The amount of entropy that is to be measured is transferred to the ice-water mixture. The amount of the melting ice is a measure of the entropy that has been supplied.

Since 1 g of liquid water contains 1.40 J/K more entropy than 1 g of ice, the amount of the entropy that is added to the flask can easily be determined. Moreover, the liquid water has a higher density than the ice. Therefore, the entropy increase can also be read at the riser tube.

This measuring procedure is not convenient for a realistic measurement. The problem is: The entropy has to be transferred to the device without producing new entropy in the process, i. e. in a reversible process. This can be done in principle and is described in [4], but it would be difficult to realize.

Origin:

In order to account for the role of energy within the network of physical phenomena, enumerating energy forms is a means of expression which is difficult to avoid. This can be seen in a citation of *F. Mohr* (1837) from the time before the discovery of the conservation of energy: "In addition to the 54 known chemical elements there exists in nature yet another agent, the name of which is Force: Under appropriate circumstances, it appears as movement, chemical affinity, cohesion, electricity, light, heat and magnetism, and from each of these forms of appearance, all of the others can be brought into being."

Disposal:

We save many words if we refrain from useless differentiations. It is often comfortable to speak about bottle milk and carton milk. It is completely useless, however, to call the process of transferring or drinking it "milk conversion," or to define the content of a glass or of the stomach as

Friedrich Herrmann

4.36 Increase of entropy when mixing pepper and salt

Subject:

"Entropy can change without creating or annihilating heat (pepper and salt)" [1]

Deficiencies:

1. First a linguistic problem: If something is created, it was not there before and is there after; if something is annihilated it was there before and is no longer there after. For this reason the quantity Q, called heat, can neither be created nor can it be annihilated. That is true at least if we use the terms "create" and "annihilate" in the current sense of the words. I believe that the above sentence is not simply a gaffe of the author, but it is a laxness that is widespread in physics. One must not be surprised, when students, and a fortiori grammar school pupils, have problems in dealing with the so-called process quantity or process function Q.

2. In statistical physics entropy is defined by

$$S = -k\sum_i p_i \ln p_i \qquad (1)$$

The formula can be applied to any discrete random variable X. p_i is the probability for X to have the value X_i. The unspecificness is on the one hand the reason for the elegance of this definition; on the other hand it leads to misunderstandings. In order to calculate an entropy value, nothing is needed but a probability distribution. All we have to know is what are different states; we do not need to know, in what the states differ, and we do not need to know by how much they differ. Entropy can be considered as one among several other quantities that describe statistical distributions, like the average value, the dispersion or the higher moments. As a consequence, definition (1) can be applied to systems and situations, which have not much to do with thermodynamics.

It means in particular that we can calculate an entropy value for systems, that are not in thermodynamical equilibrium, i.e. systems for which neither a temperature nor a chemical potential is defined. If such an equilibrium cannot establish for principal reasons, the quantity of equation (1) looses its thermodynamical meaning. Despite this fact, examples of this type are often discussed in the context of thermodynamics. So, one considers the entropy increase when shuffling cards [2] or, as in our quote, when mixing pepper and salt. In both cases no temperature and no chemical potential can be defined, and even after waiting an arbitrary long time and even with any arbitrary thermal activation no state will establish in which these quantities exist. The entropy that one has calculated has not a greater meaning than that which one would obtain by applying equation (1) on the grading of a physics test at school. In this case calculating an entropy is not more than an academic gimmick.

3. We suppose that the pepper and the salt were mentioned in order to show that temperature does not rise even though the entropy has increased during the process of mixing or shuffling. Normally, this is shown with the Joule-Gay-Lussac experiment. A gas that is under high pressure is allowed to expand into the vacuum. In the process the entropy increases by ΔS. If this amount ΔS would have been supplied, without a simultaneous increase of the volume, the temperature increase would be clearly visible. In the Joule-Gay-Lussac experiment no temperature increase is seen.

With the pepper-and-salt experiment, however, it can even not been decided if such a temperature change takes place or not, since the entropy increase that one might expect, if there was not the argument discussed above against it, would only be about 10^{-23} times the initial entropy [3]. Thus, the experiment cannot decide wether there is a temperature increase or not.

Origin:

When teaching entropy, usually no intuitive idea is associated with entropy as a macroscopic quantity. Therefore, one clings to the statistical interpretation. Then it is suggestive to make the calculation with a system for which the probability distribution is well-known or easy to obtain: dice, card games or, as in our case: pepper and salt.

Disposal:

Just as other physical quantities introduce entropy as a measure for a certain property of a body or physical system. Just as mass measures inertia or momentum measures what we usually call momentum or impetus, entropy measures what we perceive as the amount of heat that a body contains. In this way one comes without detours to a sound understanding of processes and phenomena of the everyday life and of technical devices. It is more important to learn, that the amount of entropy remains constant when the vapor passes through a turbine, than that the entropy increases by 10^{-23} when mixing pepper and salt.

The microscopic interpretation of the entropy can be done later, just as that of the temperature, the electric resistance and the mass.

Friedrich Herrmann

[1] The sentence is from a presentation (slide no. 9), that is published at the web site of the German Physical Society, and which apparently represents the opinion of the authors of the report on the Karlsruhe Physics Course.

http://www.dpg-physik.de/veroeffentlichung/stellungnahmen_gutachter/vortrag-meier.pdf

[2] *D. Meschede*, Gerthsen Physik, 21. Auflage, Springer Berlin, S. 244.

[3] *F. Herrmann, G. Bruno Schmid:* An analogy between information and energy, Eur. J. Phys. 7, 174-176 (1986)

4.37 Heat, energy and enthalpy of vaporization

Subject:

A. “For the phase transformation liquid → gaseous a certain (temperature dependent) amount of heat of transformation Q_{lg} is required.”

B. “For the vaporization of a liquid the enthalpy of vaporization is required. Heat is withdrawn from the environment or the liquid, respectively.”

C. “The heat energy that is necessary to bring a body from the liquid into the vapor phase is called latent heat of vaporization. […] Instead of latent heat one also applies the term enthalpy of transformation.”

D. “Since for a vaporization work has to be done against the molecular attractive forces, heat is consumed. The amount of heat that has to be supplied to a mass of 1 g of a liquid, in order to vaporize it at constant temperature, is called specific heat of vaporization λ. […] The heat of vaporization consists of an inner and an outer part. The outer part is spent to expand the initial volume […] to the volume of 1 g of the vapor.”

E. “In the same way energy is necessary to dissociate the particles of a body as the transition liquid → gaseous takes place. The cause of the energy requirement for the melting or evaporation process are the attractive electric forces that exist between the particles of the matter. The energy that is added in the process is then stored in the matter, the so called thermodynamic system, as potential energy of the particles.”

F. “If water is evaporating, the water molecules will stray increasingly further apart. In the process they have to move against the attractive forces that are acting between them, and they also have to push the air away. The energy that is required to do so usually comes from a heating device […] The additional energy is stored in the water vapor. In the process of condensation it is transferred to other bodies […].”

Deficiencies:

Although the authors express themselves in rather different terms, each of the quotes suffers from the fact that the process is described by a quantity that is not appropriate: the energy. The quotes show in different ways the resulting inconsistencies.

1. Quote A says: “For the phase transformation liquid → gaseous a certain (temperature dependent) amount of heat of transformation Q_{lg} is required“, “because”, one would like to continue, “the gas contains more heat than the liquid.” But that would not be true.

One would believe it is true, because the sentence of quote A translated into another context would allow for such a conclusion.

In order to paint a wall a certain amount of paint is required. Obviously the paint is first in a bucket and then on the wall.

However, in the case of the phase transition, things are different for two reasons:

– For principal reasons heat is not anywhere, it is a so called process quantity. It is hardly possible to handle the quantity verbally in a way that is appropriate to its mathematical properties and that allows at the same for an intuitive picture of the corresponding process.

– The energy that is supplied “in the form of heat” is, after being supplied, not located where one might expect to find it. Only a part of it is localized within the vapor. The remainder goes into a system that has not much to do with the phase transition: into the gravitational field.

(For those readers, who have forgotten it and those who never have learned it: the supplied energy goes partly into the vapor, since 1 kg water vapor contains more energy than 1 kg liquid water of the same temperature. But that is not all the energy. The other part is needed “to lift the atmosphere”. The vapor needs more space than the liquid. This latter part ends in the gravitational field.

2. Quote B is correct, one might believe: the supplied energy is not equal to the increase of the energy of the evaporating substance. It is equal to the increase of its enthalpy. But can that be understood? What does “required enthalpy” mean? The wording suggests that this enthalpy has to be taken from somewhere else. This, however, is not correct. What is taken from somewhere else in not enthalpy but energy.

3. The author of quote C believes he can help the reader by explaining that heat and enthalpy of vaporization are only two names for the same thing, what is not correct. The whole effort that one might have invested in explaining to the students the meaning of a process quantity is foiled.

4. According to citation D the heat of vaporization consists of two parts: an “internal” and an “external” one. If something consists of two parts it should be possible to recognize these parts in some way. But that is not the case here. If we use the tap water for drinking and cooking on the one hand, and for rinsing and washing on the other, we will not say that the tap water consists of two parts.

5. Schoolbook authors know that we cannot expose our students to statements as those cited above. (And the reason is not that the students are not intelligent enough.) That is why in schoolbooks one tries to remain on safe territory, quotes E and F: Nothing is said about the process quantities heat and work, nothing about the Legendre transform enthalpy. One only speaks about the well-behaved state variable energy. So one can use a language that can be understood by every student. However, the problem is not away. The fact that a part of the energy goes into the gravitational field is simply omitted, citation E. One might excuse this negligence with the argument, that this is only an unwanted side-effect.

In quote F this effect is mentioned. In our opinion this is the best way to deal with the subject. However, even here, something does not run smoothly. Let me try to illustrate it with a parable.

One wants to compare how much steel is needed for the construction of various suspension bridges. The bridges have the same length, carrying capacity etc., but they are designed differently. As a measure of the property in which one is interested one chooses the money that the bridge had cost. We suppose that most of this money was needed for buying the steel. But of course, money was also spent for other purposes, in particular for paying the construction company. Obviously to evaluate what was desired initially, one had taken an inappropriate measure: the monetary value instead of the mass of the steel. But this is the same mistake as that which has been committed when describing the process of vaporization. Since one wants to avoid the appropriate measure, i.e. the entropy, an “ersatz” quantity is used, the energy, which however is only partly characteristic for what one is interested in.

Origin:

1. One deals with the energy as it was done at the end of the 19th century. The following evolution has been ignored. Only at the turn of the 20th century one got able to establish local balances of the energy.

2. The balance is made for the wrong quantity. With the entropy everything would have been easy.

Disposal:

If one wants to describe the vaporization (of water for instance) by means of the energy:

Energy is supplied to the portion of water that is to be vaporized. After the process this energy is found partly in the vapor, partly as so-called potential energy in the atmosphere, which had to be lifted. (But it would be better to identify the gravitational field as the energy storage system.) The whole energy can be expressed by variables of the system “vaporizing water”: $E + pV$. That is why one can describe the process as follows: A water portion is supplied with energy ΔE. The change of the enthalpy ΔH is equal to the supplied energy ΔE. The remaining energy has been given away to the atmosphere.

The description is simper and clearer, however, when the accounting is made for the entropy. The entropy behaves corresponding to our common sense. When supplying entropy to the water portion it vaporizes. The added entropy is now found in the vapor.

Measuring the vaporization entropy is not more difficult than measuring the vaporization energy.

Friedrich Herrmann

[1] *G. Mie:* Entwurf einer allgemeinen Theorie der Energieübertragung, Sitzungsberichte der Kaiserlichen Akademie der Wissenschaften. CVII. Band VIII. Heft (1898), S. 1113

[2] *F. Jaumann:* Geschlossenes System physikalischer und chemischer Differentialgesetze, Wiener Berichte CXX, Abt. IIa, S. 385-530.

[3] *F. Herrmann:* Altlasten der Physik, Teil 1, Artikel 16: Die Messung der Entropie

4.38 The second law

Subject:

Students come across the second law in various formulations, such as:

- *Heat can never pass from a colder to a warmer body without some other change, connected therewith, occurring at the same time.*
- *It is impossible to construct an engine which will work in a complete cycle, and produce no effect except the raising of a weight and cooling of a heat reservoir.*
- *There exist irreversible processes.*

Deficiencies:

The second law makes a simple statement about entropy: Entropy can be created but not destroyed. When formulated in this way we will call it in the following the *entropy law*. It belongs to a series of other statements about the conservation or non-conservation of a substance-like quantity. However, it is rarely formulated in this way. Instead it is mostly enunciated without mentioning the physical quantity entropy. But how can this be done? By describing consequences of the asymmetrical behavior of the entropy.

This way of dealing with the key message of the second law is not convenient.

Let us discuss some formulations, that are taken from well-known textbooks, but some of which go back to the work of the great thermodynamicists of the end of the 19th century and the beginning of the 20th century.

1. Clausius, to whom the invention or introduction of the entropy is attributed, formulated the second law in several ways, among others as follows:

> *Heat cannot go by itself from a cooler to a hotter body.* [1]

Similar formulations can be found in modern textbooks, for example:

> *A process, in which finally nothing happens than taking heat energy from a colder reservoir and supplying the same amount of heat energy to a warmer reservoir, is impossible.*

Or:

> *Heat flows by itself always from the hotter to the colder body, never in the other direction.*

If we do not deny ourselves to employ the concept of entropy, we can describe the forbidden processes, to which the above statements refer, in the following way.

Because of the relation

$$dQ = TdS \qquad (1)$$

each heat current is linked to an entropy current. Therefore, we can also say: *Entropy* flows by itself from a hotter to a colder body. In this process additional entropy is produced. Thus, in the reverse process entropy would have to be destroyed. But that is forbidden by the entropy law.

We can also note that the statement is equivalent to the following:

> *Water flows by itself only downhill, never uphill.*

Also this statement is a consequence of the second law and could be put forward as a formulation of it. The fact that we consider it as trivial shows that the consequences of the second law are ubiquitous in our every-day life. Also the fact that heat goes from hot to cold and not in the revers direction is no news for whom begins to study thermodynamics.

If these statements or observations are discussed in the context of the second law, it might be appropriate to mention other phenomena in which a dissipative current is flowing: an electric current flows from high to low electric potential, a chemical reaction runs in such a way that the chemical potential of the reactants is higher than that of the products, in a frictional process momentum goes from a body of higher to one of lower velocity, etc.

2. Also Planck gives several formulations of the second law, among them:

> *It is impossible to construct a cyclically working machine, that does nothing but lift a load and cooling a heat reservoir.* [2]

Also this form of the second law can be found in modern textbooks, for instance:

> *There is no periodically working machine, that causes nothing but the production of mechanical work and the cooling of a heat container.*

Or:

> *It is impossible to construct a cyclically working machine, that causes no other effect than extract heat from a single reservoir and realize an equivalent amount of work.*

Because of equation (1) these statements are equivalent to saying that entropy cannot be destroyed. However, it is not said, that it can be created. Thus, these formulations are equivalent to only half of the entropy law: Entropy cannot be destroyed. As a consequence, the content of this kind of statement is analogue to the following:

> *There is no periodically working machine, that causes nothing but the production of mechanical work and the discharge of an electrically charged body.*

Obviously such a hypothetical machine cannot work because electric charge cannot be destroyed.

3. The aspiration to formulate the second law without mentioning entropy bears weird fruits. In the textbook by Meschede and Gerthsen [3] the second law is reduced to the simple statement:

> *There exist irreversible processes.*

The sentence is highlighted in the text. It pronounces a fact that everybody knows, even if he or she had never experienced a physics lesson. It does not allow for any conclusion about the physical cause of the irreversibility. As a physicist one could at best conclude that any of the extensive physical quantities can either be produced and not destroyed (like entropy), or be destroyed and not produced.

If one is satisfied with such a formulation one might add yet another theorem with the same explanatory power:

> *There exist reversible processes.*

From this statement we would conclude, that there exist one or more conserved quantities, but we would not know which they are.

4. It is remarkable that the second law is often formulated in different ways in one and the same textbook. Several consequences are presented as alternative formulations of the law. This seems to show that the author is not really satisfied with one single formulation. In the textbook of Macke [4] five different formulations are resumed in a table. Nobody would have the idea of formulating the conservation of the electric charge in five different manners, what would be perfectly possible.

Origin:

At the beginning the second law could not be formulated in the simple modern form, since it was not clear, that the entropy as introduced by Clausius belongs to a class of physical quantities with certain particularly simple properties: the substance-like quantities. To each of these quantities a density, a current intensity and a current density, and if applicable, a production rate can be defined. To this class of quantities belong electric charge, momentum, mass and others. The fact, that entropy is also a substance-like quantity became clear only little by little.

That is why initially it was difficult to formulate the law in an easily comprehensible way. Clausius himself proposed several versions of the law [5]:

> *The algebraic sum of all transformations appearing in a cyclic process can only be positive.*

or [6]:

> *If we call equivalent two transformations that can replace mutually without causing any remaining change, then the produced amount of heat Q of the temperature t from work has the equivalence worth* $\frac{Q}{T}$,
>
> *and the transition of the amount of heat Q of the temperature t_1 to the temperature t_2 the equivalence value*
>
> $$Q\left(\frac{1}{T_2}-\frac{1}{T_1}\right),$$
>
> *where T is a temperature function that is independent of the kind of process, by which the transformation is taking place.*

or [7]

> *When one divides the heat element by the absolute temperature that belongs to it, and integrates the resulting differential expression for the whole cyclic process, for the integral that is formed in this way the following relation is valid:*
>
> $$\int\frac{dQ}{T}\leq 0$$
>
> *where the equal sign has to be applied in those cases where all changes, of which the cyclic process consists, happen in a reversible manner, whereas in the cases where the changes happen in a non reversible way, the sign < is valid.*

And, as mentioned above, Clausius also formulates:

> *Heat cannot go by itself from a cooler to a hotter body.* [1]

Without further explanations, the first three quotes are nowadays hardly comprehensible. This explains why the second law was considered a difficult matter.

From Clausius' work it is hardly possible to deduce that entropy is a substance-like quantity. The literature of that time mainly treats the consequences of the new "insight" that heat is a form of energy for the work of Carnot. According to Carnot the amount of "heat" that enters a thermal engine is equal to the amount that leaves it. If one interprets heat as energy, the amount of "heat" that enters the engine is greater than the amount that leaves it. Therefore it was concluded that Carnot was not right, since there was the conviction that heat is something existing in nature and that its true nature, i.e. its being a form of energy, had now been discovered. Under this premise Carnot would indeed be wrong. Only slowly it became clear that Carnot was not wrong. There were simply two different concepts of heat. Carnot's heat was another physical quantity than the new heat energy. It turned out that Carnot's heat coincided with the newly introduced entropy, and that meant that the quantity introduced by Clausius was actually not new. In 1908 Ostwald says it in clear words [8]:

> *That thermodynamic quantity which could be compared with an amount of water, has not entered into the general awareness. It received the scientific name of entropy and plays a role in the theory of thermal phenomena that corresponds to its significance. However, it has not yet entered into the schools and thus into the knowledge of the average persons, and therefore we have to limit ourselves to say, that it is really comparable with an amount of water, since its amount does not change when going through a (ideal) machine.*

Here, Ostwald alludes to Carnot's comparison of a heat engine with a water wheel.

Three years later, in 1911 Callendar also showed that Carnot's calorique has to be identified with the entropy and that Clausius' introduction of the entropy is unnecessarily complicated [9,10].

In the same year, Jaumann [11] has published an article in which he gives the first formulation of a local balance of the entropy. He introduces an entropy flow, an entropy density, an entropy flow density and a production rate. Now it was possible to formulate the second law in a way that is analogue to the current formulation of the balance laws for other substance-like quantities: energy, momentum and electric charge:

Energy cannot be produced or destroyed; momentum cannot be produced or destroyed; electric charge cannot be produced or destroyed.

However, only few textbooks, formulate the second law in this simple way. An example is Grimsehl [12]:

> *In a closed system all processes happen in such a way that the entropy never decreases.*

Or in Joos [13]

> *All changes of state that occur in a closed system are such that the entropy increases.*

Both these books are already rather advanced in years. We can conclude that an old teaching tradition cannot be eradicated by the insight of a few persons. More than a hundred years after Ostwald, entropy „has not yet entered [...] into the knowledge of the average persons".

Disposal:

The disposal is particularly simple: Introduce entropy in the sense of Carnot's calorique as a measure of an amount of heat. Then the second law tells us, that this „heat" can be produced but not destroyed, a statement that each layman can confirm based on his everyday experience.

Friedrich Herrmann

[1] *R. Clausius:* Zur Geschichte der mechanischen Wärmetheorie (About the history of the mechanical theory of heat), Annalen der Physik, Vol. 221, Heft 1 (1872), S. 132

[2] *M. Planck:* Thermodynamik, Verlag von Veit & Comp., Leipzig (1897), S. 80

[3] Meschede, Gerthsen Physik, 21. Auflage, Springer, Berlin (2002), S. 248

[4] *W. Macke:* Thermodynamik und Statistik (Thermodynamics and statistics), Akademische Verlagsgesellschaft, Geest & Portig K.-G., Leipzig (1962), S. 120

[5] *R. Clausius:* Über eine veränderte Form des zweiten Hauptsatzes der mechanischen Wärmetheorie, Annalen der Physik und Chemie (About a modified form of the second law of the mechanical theory of heat), Band XCIII (1854), S. 504

[6] *R. Clausius:* Abhandlungen über die mechanische Wärmetheorie, Erste Abtheilung (Treatises about the mechanical theory of heat, first part), Druck und Verlag von Friedrich Vieweg und Sohn, Braunschweig (1864), S. 143

[7] Clausius, Abhandlungen über die mechanische Wärmetheorie, Zweite Abtheilung (Treatises about the mechanical theory of heat, second part), Druck und Verlag von Friedrich Vieweg und Sohn, Braunschweig (1867), S. 3

[8] *W. Ostwald:* Die Energie (The energy), Verlag von Johann Ambrosius Barth, Leipzig (1908), S. 77

[9] *H. L. Callendar:* The caloric theory of heat and Carnot's principle, Proc. Phys. Soc. London **23** (1911), S. 153

[10] *H. L. Callendar:* Science, Vol. XXXVI, No. 924 (1912), S. 321

[11] *F. Jaumann:* Geschlossenes System physikalischer und chemischer Differentialgesetze (Complete system of physical and chemical differential laws), Wiener Berichte CXX, Abt. IIa (1911), S. 385-530

[12] *W. Schallreuter:* Grimsehl, Lehrbuch der Physik, B. G. Teubner Verlagsgesellschaft, Leipzig (1957), S. 467

[13] *G. Joos:* Lehrbuch der Theoretischen Physik, Akademische Verlagsgesellschaft, Frankfurt am Main (1959), S. 488

4.39 The adiabatic state equations

Subject:

A reversible adiabatic process can be described by means of the following equations, sometimes called *adiabatic state equations*:

$$p \cdot V^{\gamma} = \text{const}$$
$$T \cdot V^{\gamma-1} = \text{const}$$
$$T \cdot p^{\frac{1}{\gamma}-1} = \text{const}$$

Here, γ is the ratio of the specific heat capacities at constant pressure and that at constant volume:

$$\gamma = c_p/c_V \, .$$

Deficiencies:

In principle, there is nothing to object. However, what idea of these processes is created in the mind of the students?

1. For an ideal gas with a constant amount of substance, i.e. for a typical thermodynamic system, the Gibbs fundamental equation reads

$$dE = TdS - pdV \, . \qquad (1)$$

It expresses, among other things, the fact that the system has two degrees of freedom.

One is well-advised not to change the values of various variables at the same time. Therefore, one likes to consider processes in which one of the five variables in equation (1) is held constant. In this way the system is reduced to a system with only one degree of freedom. The corresponding processes are given proper names: isothermal, isobaric, isochoric, isoenergetic and…? Adiabatic!

As is well-known, “adiabatic” means “impassible”. Thus, instead of referring to the variable that is held constant, i. e. the entropy, instead of calling the process isentropic, the name tells us what we have to do in order to keep the entropy constant: the walls of the container of the gas have to be impenetrable for heat, and thus for entropy. Don’t object, that isentropic means something different from adiabatic. Of course, an adiabatic process can be realized in which entropy is produced, and which is, as a consequence, not isentropic; it is also possible to realize a process that is isentropic although the walls of the container are not impenetrable for heat. These remarks are correct, but the common use of the term is for processes that are adiabatic and isentropic at the same time.

2. One can define and measure quite a lot of coefficients that characterize a system. In the case of a system that corresponds to equation (1) there are

- the compressibility at constant temperature;
- the compressibility at constant entropy;
- the volumetric thermal expansion coefficient;
- the pressure coefficient;
- the specific heat capacity at constant volume;
- the specific heat capacity at constant pressure;
- and others that do not have proper names.

However, only three of these coefficients are independent; the remaining coefficients can be calculated from them. For the purpose of teaching (at school and university) we are not only confronted with the question of which of these coefficients should be introduced but also how to introduce them. When we ask for the meaning of the “adiabatic index” γ , the expected answer is: the ratio of the specific heats at constant pressure and at constant volume. It is not easy to get an idea of the meaning of c_p and c_V separately. But what is the meaning of the ratio of these coefficients?

3. There are two extreme ways of disposing of the thermal variables: either the temperature is held constant, by allowing for a perfect entropy exchange with the ambient, or the entropy is held constant by preventing an entropy exchange with the surroundings [1]. These two conditions are complementary; they are equally important and should be treated on an equal footing. We can say that processes in small systems tend to be isothermal; in large systems they are isentropic. The smaller the system, the “more isothermal” are the processes. And we also have: The slower the process, the “more isothermal” it is, the faster the process, the “more isentropic”. Thus we have the rule: small and slow → isothermal; large and fast → isentropic. Or: „Small fishes are isothermal, big fishes are isentropic.” Weather phenomena are large-scale phenomena. Therefore they are essentially isentropic.

4. That adiabatic state equation which usually is considered in the first place, see above, is the most uninteresting of the three, since the p-V relationship is very similar in an isothermal and an isentropic process. More interesting are the second and the third equation. Let us write them in a more convenient way. By using

$$\alpha = \frac{1}{\gamma - 1}$$

we get

$$V \cdot T^{\alpha} = \text{const} \qquad (2)$$

and

$$p \cdot T^{-(1+\alpha)} = \text{const} \qquad (3)$$

Equation (2) tells us that and how the temperature of a gas decreases upon expansion, and equation (3) tell us that the air at high altitudes, where the pressure is low, is cold. Some values of α are listed in the table.

	α
air	2,5
water vapor	3,3
carbon dioxide	3,4
helium	1,5
light	3

In our natural and technical environment the isentropic processes are more important. They are those processes that we try to realize in thermal engines, and they are the relevant processes in weather phenomena.

5. When defining the exponent of equation (2) by means of the heat capacities, the behavior of a gas in an isentropic compression or expansion appears unintuitive. However, it is not difficult to present such a process as a very natural phenomenon without referring to the heat capacities: When a gas is compressed, also “the entropy is compressed”, i.e. its density is increased. Then it is normal to expect that the corresponding intensive quantity, the temperature, also increases.

Origin:

The wide-spread aversion against entropy also manifests itself in the treatment of isentropic processes.

Disposal:

Introduce the equations

$$V \cdot T^{\alpha} = \text{const}$$

and

$$p \cdot T^{-(1+\alpha)} = \text{const}$$

Do not define the exponents via the specific heats. The equations tell us how the temperature reacts upon a change of the volume and of the pressure.

Friedrich Herrmann

4.40 The barometric formula

Subject:

When discussing the atmospheric pressure we treat the barometric formula. Thereby we advert to the fact that the underlying assumption that the temperature is independent of the height is not realistic.

1. “Equation (…) allows to determine the height difference from the air pressure at two different altitudes (barometric height measurement). In reality the condition of constant temperature is not fulfilled. For height differences that are not too great the mean value of the temperatures at the altitudes h_0 and h can be used.”

2. “This equation, which is usually called the barometric formula, opens the possibility to calculate the difference in altitude between two points, if the pressure and the temperature of the air is known at both stations. The two equations (…) are valid for an isothermal atmosphere; in nature normally temperature changes with the altitude. However, the formula for the isothermal atmosphere can be applied without causing a greater error, if for T the average value of the temperature between the two levels is introduced into the formula.”

3. “Under the (unrealistic) condition that the atmosphere of the Earth has a unique temperature T a formula for the dependence of the gas pressure p on the height h can be given: …”

4. “However, here we have assumed a constant temperature, since we have used Boyle’s law; a height formula that is based on the adiabatic law p/ρ^{κ} = const, that is derived in an analogous way is more appropriate.”

5. “In reality the temperature within the troposphere (until 10-12 km) decreases in general with the altitude. The troposphere for dry air is better described with an adiabatic-indifferent stratification.”

Deficiencies:

First, the barometric formula is derived under the assumption that the troposphere is in thermal equilibrium in a vertical direction, for instance in the form

$$p = p_0 e^{-\frac{Mgh}{RT}}$$

and then comes the disclaimer.

Sometimes (quotes 1 and 2) it is proposed, how one can take profit of the formula despite of this shortcoming: by applying it to small intervals of the altitude and using the average temperature.

By treating the problem mathematically the impression of rigor is created. But then it is admitted, that the premises of the calculation are not fulfilled “in reality”, that they are “unrealistic”. Sometimes it is even declared how one could have done better, quotes 4 and 5.

One might be inclined to justify this procedure, by arguing that it is an idealization. Isn’t it like in mechanics when we neglect friction? It is not. When deriving the barometric formula one does not neglect a small perturbing effect. Regarding the entropy exchange in a vertical direction the opposite of what happens in reality is assumed.

Two extreme cases of thermodynamic processes can be distinguished: isothermal and isentropic processes. In the case of the troposphere the isentropic behavior, i.e. the assumption of a constant molar entropy is a good approximation; the assumption that temperatures equalize is a bad one. “Isothermal” is not an approximation of “isentropic”, it is the opposite of it [1].

It is obvious that the assumption of a height-independent temperature is a bad approximation. It presupposes that the air at high altitudes is in thermal equilibrium with that at low altitudes. However, thermal equilibrium can establish only if a non-convective entropy flow is possible: an entropy exchange between one portion of air with another. A movement of the air, and also a strong mixing by turbulence cannot establish thermal equilibrium between different altitudes. On the contrary: a strong mixing of the air is the condition for the establishment of the actual, natural temperature gradient.

Apparently, the natural temperature gradient is not taken as seriously as the pressure gradient. Who would have the idea to calculate the temperature gradient and doing so admit that the pressure gradient is zero?

Two mechanisms exist that contribute to an equalization of the temperature: When water evaporates at low altitude or on the ground, condenses then in high altitude and goes back down as rain, thereby traversing the air we have an entropy transport in the upward direction with the tendency to reduce the temperature difference. A second effect is the heat exchange by radiation that acts in the same direction.

These are additional effects, which one will neglect at the beginning when trying to understand the working of the troposphere. One begins with the dry-isentropic (= dry-adiabatic) atmosphere.

The fact that the barometric formula is simple and can easily be taught to a beginner, cannot be a justification of the assumption of a constant temperature either, since the formula for the pressure gradient in the dry-isentropic atmosphere is not more complicated: it is a power function (with a fractional exponent):

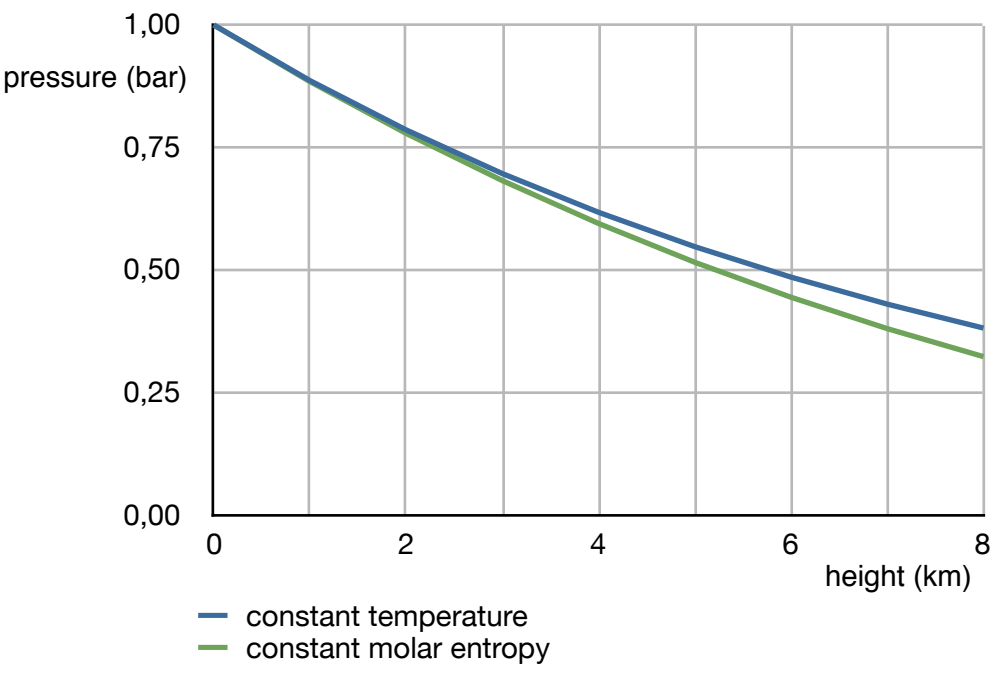
$$p(h) = p(0)\left(1 - \frac{Mg}{c_p T(0)} h\right)^{\frac{c_p}{R}}$$

The temperature gradient, that is supposed to not exist, even obeys a simpler law. It is a linear relationship[1]:

$$T(h) = T(0) - \frac{Mg}{c_p} h$$

Another peculiarity that that follows from quotes 1 and 2:

One tries to explain how the altitude can be calculated from the pressure. Since the barometric formula does not work very well, it is proposed to limit to small height intervals.



The figure shows the pressure as a function of height, as it follows from the barometric formula and as it is for the dry-isentropic atmosphere. One can see: When the formula is applied only to small intervals, it indeed doesn’t matter if one takes one or the other formula. However one can save the effort: a linear approximation does just as well.

Origin:

1. If one is at loggerheads with the entropy the condition of constant entropy may appear as a difficulty. And indeed it is difficult to formulate it if one wants to avoid the concept [2].

2. The barometric formula is welcome because it seems to be an example for the acting of the Boltzmann formula.

3. The formula can be derived with only mechanical arguments. In this way one avoids to have to deal with the unloved thermodynamics. This at least is the impression. From thermodynamics one only overtakes Boyle’s law, which only contain the mechanical quantities p and V. Thereby one does not really notice, that the mechanical and thermal variables are strongly coupled. Keeping the temperature constant seems to be a measure of the same kind as keeping it constant when verifying Ohm’s law.

Disposal:

1. Make clear, that in the atmosphere which is in equilibrium there is not only a natural pressure gradient but also a natural temperature gradient, and that this temperature gradient cannot be considered a harmless deviation from the thermal equilibrium that has not yet established.

2. Treat the idealization of the dry-isentropic atmosphere. Here the temperature curve is particularly simple, namely linear.

Friedrich Herrmann

[1] A consequence of this law would be that the temperature would attain the absolute zero at a height of 30 km approximately. However, long before the air would have become liquid and the conditions for the application of the formula are no longer fulfilled.

[1] G. Job, *Die Temperaturschichtung der Atmosphäre*, Altlasten der Physik, Aulis Verlag Deubner (2002), Köln, S. 117

[2] Altlasten der Physik 164

4.41 No temperature – no entropy?

Subject:

Recently I read (in a context that is not interesting here):

"The system has entropy, therefore it has a temperature."

The sentence did not mean that the temperature of the system has a value that is greater than zero, but that the quantity temperature has a value at all.

Deficiencies:

First, a general remark about the "has" in the context of a physical quantity. If we say that a particle has no electric charge, that the photon has no rest mass, or that a car has no momentum, we always mean that the value of the corresponding physical quantity is zero, i.e. $Q = 0$ C, $m_0 = 0$ kg or $p = 0$ kg · m/s. Something else is meant when we say a system has no temperature. It does not mean $T = 0$ K. (The same applies to the chemical potential.) It rather means that its state can not be described by a temperature. In other words, the system is not in a state of thermodynamic equilibrium, or the occupation distribution of the microstates does not correspond to any of the known statistical functions.

If this is meant by the sentence cited above, the statement is not necessarily correct. It cannot be concluded from $S > 0$ that the considered system has a temperature. It only has a temperature if it is in thermodynamic equilibrium: if all "accessible microstates" are occupied with the same probability. The entropy then simply calculates as

$$S = k \ln W \tag{1}$$

In general, however,

$$S = -k \sum_{i=1}^{n} p_i \ln p_i \tag{2}$$

Equation (1) follows from equation (2) if all probabilities p_i are equal:

$$p_1 = p_2 = p_3 = p_4 = \ldots = p_n = 1/W .$$

But is not everything that surrounds us in good approximation in thermodynamic equilibrium? Doesn't upon a change of an external parameter the thermodynamic equilibrium establish so fast that nonequilibrium states do not matter?

Not at all. That the equilibrium does not establish can have two causes:

First, the density of the interacting particles is too low. An example is the atmosphere of the Earth at high altitude.

Second, the density is high, but the particles do not interact with each other. This phenomenon is omnipresent: light that has not already been generated in a state of thermodynamic equilibrium will later have no chance to reach this state, unless it gets help – such as the famous carbon particle of Max Planck.

An example of light that is not in equilibrium is the light we get from the Sun. A priori the conditions for obtaining light in equilibrium are favorable: Its source is in good approximation a black body. The frequency distribution of the sunlight that arrives here at the Earth corresponds, to a good approximation, to that of light in thermodynamic equilibrium. What makes the difference from radiation in thermal equilibrium is the angular distribution. For the light to be in thermodynamic equilibrium, it would have to be distributed isotropically, but from that it is far away. So it happens that the relation between energy current and entropy current for the light from the Sun is not

$$P = T \cdot I_S .$$

as one would expect for a transport with light in thermodynamic equilibrium. We rather have

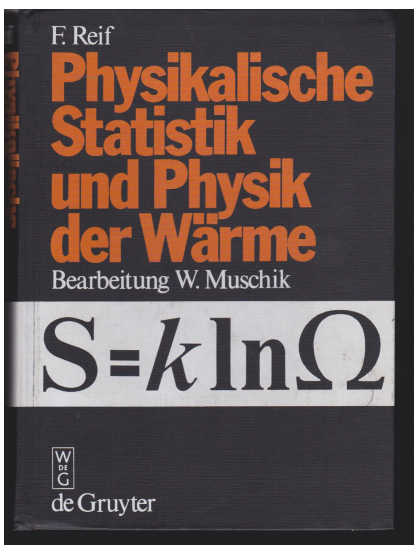
$$P = (3/4)\ T \cdot I_S$$

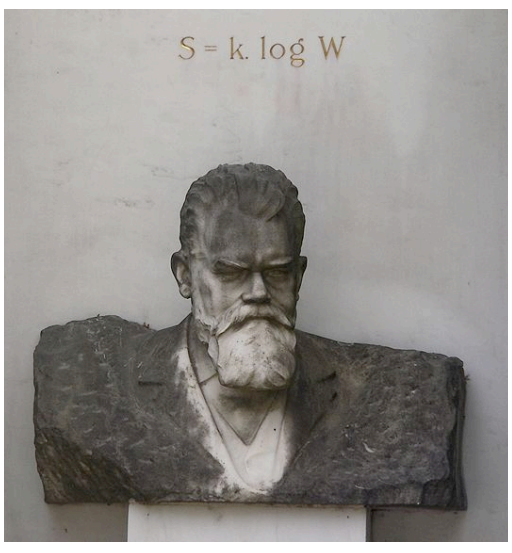
where, however, T is not the temperature of the light that we receive, but that of the surface of the Sun.

Origin:

Probably the fact that the equation $S = k \ln W$ (or $S = k \ln \Omega$) has become so emblematic of entropy, Fig. 1. (Similarly to Einstein's equation $E = mc^2$.)

Equation (2) is perceived more as a curiosity, or as a measure of data.





Disposal:

1. I recommend to handle the language carefully when referring to the values of physical quantities. About an extensive ("substancelike") quantity one can and should speak in the same way as when speaking about a substance: a system has much or little or no entropy (or electrical charge, mass, momentum). The linguistic environment of intense quantities is quite different: a temperature (an electric potential, a velocity) is high or low, and in the case of temperature and chemical potential, a system may not have these quantities.

2. Introduce formula (2) for the entropy before dealing with the (often realized) special case of thermodynamic equilibrium. Thus, one learns that entropy has a much broader meaning than temperature (and chemical potential that we have not discussed here).

3. One has to be prepared to meet inconsistencies if light that does not correspond to the thermodynamic equilibrium distribution is simply

assigned the temperature of the light source.

Friedrich Herrmann

4.42 The entropy of the universe

Subject:

Most of our studies were done without the cosmos. But then it finally comes: in thermodynamics - thermodynamics seems to need it. So we find the 2nd law in a textbook for the university:

> "In a reversible process, the entropy change of the universe is zero. By 'universe' we mean the totality of a system and its environment.
>
> In an irreversible process, the entropy of the universe increases.
>
> There is no process by which the entropy of the universe decreases."

Or in a textbook for the secondary school:

> "The entropy of the universe is constantly increasing or is not changing."

Or in Wikipedia under the keyword Exergonic process:

> "All physical and chemical systems in the universe follow the second law of thermodynamics and proceed in a downhill, i.e., exergonic, direction."

Deficiencies:

1.The question of what is the entropy of the universe is a difficult one, and in such a simple context as the second law it might be better not to enter the mined area of thermodynamics of the cosmos. To speak of the entropy of the universe, one would have to sum up the contributions of all parts of the universe. But how is it done? The entropy that all parts have now? Then the question is how to decide on the simultaneity of distant space-time points.

2. Suppose the size of the universe is infinite. (This idea is somewhat metaphysical, but apparently hardly anyone has a problem with it.) Then the problem arises that entropy is also infinite, and consequently must have always been infinite. Can it still increase then? Certainly it can. You only have to formulate it locally, but that means without the universe. Already in 1897 Planck [1] points out in his *Thermodynamics* that the entropy of the universe "cannot be defined".

3. Why does the universe have to serve to formulate the entropy theorem, but not the conservation of the electric charge or of momentum or the baryon number?

Why don't we formulate:

> "There is no process by which the electric charge of the universe changes."

The answer is clear: because there is a much simpler way.

4. Here, again, the idea is promoted that entropy is a particularly transcendent quantity. Entropy once again needs a special treatment..

Origin:

Already in its beginnings one had the idea to ask the question about the importance of entropy for the development of the "universe". Apparently, it was brought up by W. Thomson [2]. Clausius [3] notes 1865 in this regard:

> "… The application of this proposition [of the second law] to the whole universe leads to a conclusion which was first pointed out by W. Thomson and of which I have already mentioned in a recently published treatise. For if in all the changes of state occurring in the universe the transformations of a certain sense exceed those of the opposite sense in size, then the overall state of the universe must change more and more in that former sense, and the universe must thus approach without ceasing a limiting state."

(Clausius uses the term "transformation" for the quantity that he later baptized entropy.)

From the viewpoint of that time, these remarks seemed unproblematic, for no one could have guessed in what a difficult context the statements would be placed by the General Theory of Relativity and modern cosmology. Also the possibility of expressing the second law locally, i.e. by a continuity equation, was still far in the future. The local entropy balance was first formulated in 1911 by Jaumann [4].

Disposal:

Also in this context I recommend to take Wheeler's advice to heart: "Physics is simple only when analyzed locally".

When it comes to characterize the physical quantity entropy, it is enough to say that entropy can be created but not destroyed. Everyone understands this sentence.

If one wants it more mathematically, then one may write down the local balance equation (= continuity equation) [4]

$$\frac{\partial \rho_S}{\partial t} + \text{div}\, j_S = \sigma_S$$

(ρ_S = entropy density, j_S = entropy flux density, σ_S = density of the production rate), and notes that the production term on the right is never negative.

Friedrich Herrmann

[1] M. Planck: *Vorlesungen über Thermodynamik*, Verlag von Veit & Comp. Leipzig 1897, S. 94.

[2] W. Thomson: *On a universal tendency in nature to the dissipation of mechanical energy*, The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, Series 4, 1852, S. 306. https://www.tandfonline.com/doi/abs/10.1080/14786445208647126

> „Within a finite period of time past the earth must have been, and within a finite period of time to come the earth must again be, unfit for the habitation of man as at present constituted, unless operations have been, or are to be performed, which are impossible under the laws to which the known operations going on at present in the material world are subject."

[3] R. Clausius: *Ueber verschiedene für die Anwendung bequeme Formen der Hauptgleichungen der mechanischen Wärmetheorie*, Annalen der Physik und Chemie, Band CXXV, No. 7, S. 397-400.

[4] G. Jaumann: *Geschlossenes System physikalischer und chemi-scher Differentialgesetze*, Sitzungsber. Akad. Wiss. Wien, Mat.-Naturw. Klasse, Abt. IIA 120, 1911, S. 385-530.

4.43 Cooling with liquid nitrogen

Subject:

> "Cooling with liquefied nitrogen has a number of advantages. Low investment and operating costs are just a few of the reasons, apart from the simple and low-maintenance application. This cooling method offers further advantages not only in terms of economy, but also in terms of safety and environmental protection. Liquid nitrogen is non-combustible, non-toxic and no waste is produced."
>
> "Magnets in large accelerator facilities such as CERN in Geneva are usually cooled to almost absolute zero with liquid helium. The European infrared satellite Herschel was also brought to the lowest temperatures with helium in order to be particularly sensitive to the thermal radiation of cosmic objects."
>
> "Coolers are used in many technical devices that heat up. Mostly, passive cooling, i.e. the transfer of heat via radiators to the surrounding air, is used. The best-known example is the refrigerator for preserving food. Water cooling is usually used in motor vehicles, while air cooling is mainly used in computers. Another large field of application is, for example, air conditioning."
>
> "Processor cooling: A distinction is made between air cooling, water cooling, boiling cooling, Peltier cooling and dry ice cooling."

Deficiencies:

Cooling means bringing an object to a low temperature or keeping it at a low temperature. This is done by removing entropy from the body.

In principle, there are two possibilities:

1. There is an environment that is colder than the body to be cooled. Then one only has to make sure that the entropy follows its natural drive from warm to cold. Example: the cooling of the car engine.

2. If there is no such environment (because the temperature of the body to be cooled is below the ambient temperature), entropy must be moved from the lower temperature to the higher ambient temperature. This requires energy as well as a suitable device: the heat pump (which should better be called entropy pump).

There is a similar problem in mechanics. To bring a body (e.g. a vehicle) to a higher speed, or to keep it at a high speed, one has to provide it with momentum.

Here, too, there are two possibilities:

1. One couples the body to a system that already has a high velocity. The momentum then follows its natural drive from high to low velocity. The same thing is done when braking: the brake establishes a momentum conducting connection between the vehicle and the earth, and the momentum flows by itself into the earth.

2. If one does not have an "environment" with the desired higher velocity, the momentum must be "pumped" from the environment into the vehicle using energy. This is exactly what the car engine does.

We do not need to tell the corresponding electrical and chemical stories.

When about cooling with liquid nitrogen, helium or ice cubes is talked, one gets the impression that the so-called coolant is the decisive factor. The coils of a magnet are cooled with liquid helium. But that only means cooling something with something else that is already cold. Cooling is reduced to adjusting two temperatures.

But who cools the helium? How does helium get rid of its entropy? Nothing is said about that, at least in our quotations. It is simply liquefied.

In our third quote there is mention of the refrigerator and air conditioning. If I have understood the text correctly, the author is only concerned with how the entropy in the heat exchanger at the back of the refrigerator is passed on to the ambient air. Apparently, this is where the refrigerator is cooled. That which constitutes the refrigerator, namely its heat pump, seems to be less important.

In the last quote (from Wikipedia, but somewhat alienated) it's a funny mix-up. Peltier element and dry ice are mentioned in one breath. The first is a heat pump, the second only a cold substance from which entropy had previously been pumped out.

The decisive element for cooling, the heat pump, is either not mentioned at all or only appears as a technical detail. One needs it "only" to liquefy something, or to produce the ice cubes for the coke.

Origin:

How can one express it more clearly if one does not want to or cannot mention entropy? Accordingly, one cannot of course speak of pumping up entropy. And with the "substitute constructions" thermal energy or enthalpy it becomes complicated. It's better not to say anything at all.

Disposal:

In general, one should not reduce Carnot's work to the somewhat entangled Carnot cycle, which is only an example in his work. Rather, one adopts his ingenious idea: the comparison of a thermal engine with a water wheel. The fact that he had not yet been able to make a comparison between a heat pump and a water pump was only due to the fact that there were no heat pumps at the time.

And if one speaks of cooling, one puts the cooling machine (heat pump) into the foreground. Its function is easy to describe: The heat pump brings entropy from cold to warm under energy expenditure – just like a water pump brings water from low to high pressure. This is easy to explain. The technical tricks that are used can be explained afterwards. Or one leaves them out completely.

Finally, a suggestion: Do not differentiate between heat pumps and chillers. Of course, the devices may be built differently, but at least a hint that they do the same would be helpful for the students.

Friedrich Herrmann

4.44 The entropy of the universe

Subject:

Most of our studies were done without the cosmos. But then it finally comes: in thermodynamics - thermodynamics seems to need it. So we find the 2nd law in a textbook for the university:

> *In a reversible process, the entropy change of the universe is zero. By "universe" we mean the totality of a system and its environment.*
>
> *In an irreversible process, the entropy of the universe increases.*
>
> *There is no process by which the entropy of the universe decreases.*

Or in a textbook for the secondary school:

> *The entropy of the universe is constantly increasing or is not changing.*

Or in Wikipedia under the keyword Exergonic process:

> *All physical and chemical systems in the universe follow the second law of thermodynamics and proceed in a downhill, i.e., exergonic, direction.*

Deficiencies:

1.The question of what is the entropy of the universe is a difficult one, and in such a simple context as the second law it might be better not to go into the mined area of thermodynamics of the cosmos. To speak of the entropy of the universe, one would have to sum up the contributions of all parts of the universe. But how is it done? The entropy that all parts have now? Then the question is how to decide on the simultaneity of distant space-time points.

2. Suppose the size of the universe is infinite. (This idea is somewhat metaphysical, but apparently hardly anyone has a problem with it.) Then the problem arises that entropy is also infinite, and consequently must have always been infinite. Can it still increase then? Certainly it can. You only have to formulate it locally, but that means without the universe. Already in 1897 Planck [1] points out in his *Thermodynamics* that the entropy of the universe "cannot be defined".

3. Why does the universe have to serve to formulate the entropy theorem, but not the conservation of the electric charge or of momentum or the baryon number?

Why don't we formulate:

> *There is no process by which the electric charge of the universe changes.*

The answer is clear: because there is much simpler way.

4. Here, again, the idea is promoted that entropy is a particularly transcendent quantity. Entropy once again needs a special treatment.

Origin:

Already in its beginnings one had the idea to ask the question about the importance of entropy for the development of the "universe". Apparently, it was brought up by W. Thomson [2]. Clausius [3] notes 1865 in this regard:

> *"… The application of this proposition [of the second law] to the whole universe leads to a conclusion which was first pointed out by W. Thomson and of which I have already mentioned in a recently published treatise. For if in all the changes of state occurring in the universe the transformations of a certain sense exceed those of the opposite sense in size, then the overall state of the universe must change more and more in that former sense, and the universe must thus approach without ceasing a limiting state."*

(Clausius uses the term "transformation" for the quantity that he later baptized entropy.)

From that point of view, these remarks still seemed unproblematic, for no one could have guessed in what a difficult context the statements of the General Theory of Relativity and modern cosmology would be placed. Also the possibility of expressing the second law locally, i.e. by a continuity equation, was still far in the future. The local entropy balance was first formulated in 1911 by Jaumann [4].

Disposal:

Also in this context I recommend to take Wheeler's advice to heart: "Physics is simple only when analyzed locally".

When it comes to characterize the physical quantity entropy, it is enough to say that entropy can be created but not destroyed. Everyone understands this sentence.

If one wants it more mathematically, then one may write down the local balance equation (= continuity equation) [4]:

$$\frac{\partial \rho_S}{\partial t} + \text{div}\, j_S = \sigma_S$$

(ρ_S = entropy density, j_S = entropy flux density, σ_S = density of the production rate), and notes that the production term on the right is never negative.

[1] M. Planck: *Vorlesungen über Thermodynamik*, Verlag von Veit & Comp. Leipzig 1897, S. 94.

[2] W. Thomson: *On a universal tendency in nature to the dissipation of mechanical energy*, The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, Series 4, 1852, S. 306. https://www.tandfonline.com/doi/abs/10.1080/14786445208647126

> „Within a finite period of time past the earth must have been, and within a finite period of time to come the earth must again be, unfit for the habitation of man as at present constituted, unless operations have been, or are to be performed, which are impossible under the laws to which the known operations going on at present in the material world are subject."

[3] R. Clausius: *Ueber verschiedene für die Anwendung bequeme Formen der Hauptgleichungen der mechanischen Wärmetheorie*, Annalen der Physik und Chemie, Band CXXV, No. 7, S. 397-400.

[4] G. Jaumann: *Geschlossenes System physikalischer und chemi-scher Differentialgesetze*, Sitzungsber. Akad. Wiss. Wien, Mat.-Naturw. Klasse, Abt. IIA 120, 1911, S. 385-530.

Friedrich Herrmann

4.45 The ideal gas law and the undesired quantities entropy and chemical potential

Subject:

There seems to be consensus that the general gas equation

$$p \cdot V = n \cdot R \cdot T \qquad (1)$$

is the most important equation to describe the thermodynamics of gases. It is usually the only equation of state for gases that is treated at schools and colleges.

Equations of state describe real systems, usually matter, and matter is complicated. Fortunately, gases can be described in good approximation by a fantastically simple equation of state. There is no material constant (unless you proceed like the meteorologists, who instead of the amount of substance take the mass, so that the gas constant becomes material-dependent, a kind of masochism so to speak). And it does not only apply to the normal gases; also other systems, which are not even called gases, follow the law: substances in diluted solution.

Certainly, gases, even if they are sufficiently ideal, have individual and complicated properties, and that is why, in addition to the ideal gas law, sometimes other equations of state are needed, such as the caloric equation of state. However, we are only interested in the ideal gas law.

Deficiencies:

Let us ask an obvious question: Which are the variables of the system under consideration, i.e. the gas? One might answer: as many as one likes – because we are free to construct or define any number of variables.

However, if we ask for the variables within the framework of the established rules of physics and especially of thermodynamics, we proceed differently: We write down Gibbs' fundamental relation:

$$dE = TdS - pdV + \mu dn \qquad (2)$$

It tells us that the energy of our system can be changed in three ways, namely by changing the entropy S, the volume V or the amount of substance n (or two of them at the same time or all the three at the same time). In other words, our system has three degrees of freedom. We have decided about this by writing down equation (2). (We could also have decided differently: e.g. that we do not want to change the amount of substance of the gas. Then we would have omitted the term μdn. Or that we want to accelerate the gas. Then we would have added a term $\vec{v}d\vec{p}$.)

If we now decide for equation (2) and ask what the variables of the system are, the answer is: first the energy, and then 6 others, or 3 pairs, namely, S and T, p and V, and μ and n.

If we look at the ideal gas law, however, we find that two of the 6 variables do not appear. Where have they gone? Of course we can say: we were lucky. If they were still there, the equation would perhaps be more complicated. But most of all: we are lucky, because S and μ are just those quantities that we might not want to have too much to do with.

That they do not appear is also quite practical for the following reason: The system has three degrees of freedom, equation (1) has 4 variables. So we can calculate the fourth from three arbitrarily given ones.

However, for those who keep asking about the two variables S and μ, the question is not answered: How do S and μ behave if we change the other variables?

Of course, there are relationships that cannot be found without additional information about the system, for example: How does the temperature depend on the entropy if the volume and the amount of substance are held constant? Equation (1) tells us nothing about that. An equation of state is an equation of state. It is not a Hamiltonian, Lagrangian or Massieu-Gibbs function which contain the complete information about a system.

Nevertheless, is it not possible to say anything about the chemical potential and entropy without further information? Of course it is possible. A short calculation (using only equation (1)!) leads to

$$S(V) - S(V_0) = n \cdot R \cdot \ln\frac{V}{V_0} \qquad (3)$$

$$\mu(p) - \mu(p_0) = R \cdot T \cdot \ln\frac{p}{p_0} \qquad (4)$$

Equation (4) can also be written in the form

$$\mu(p) - \mu(p_0) = R \cdot T \cdot \ln\frac{c}{c_0} \qquad (5)$$

where c is the concentration.

The three equations apply to processes where V or p is changed but the temperature is kept constant.

Of course, no material constant enters here either.

Equation (3) tells us that at constant temperature the entropy increases with the volume. Or in other words: When a gas expands at a constant temperature, it absorbs heat.

Equation (4) tells us: The chemical potential increases (at constant temperature) with the pressure. From the equation follows directly the law of mass action. (One can also say that equation (4) is the law of mass action.) It also allows us to derive the barometric formula in a few lines without having to use force equilibria.

Equation (5) tells us among other things: The drive for the diffusion of a substance between two locations where the concentration differs by a factor of 10 is always the same, regardless of whether it goes from 0.1 to 0.01 mol/l or from 0.00001 to 0.000001 mol/l. It is responsible for the function of every electrochemical cell.

The equations (3) to (5) are also mathematically simple, and they are the expression of a rule of thumb: In an ideal gas, everything is linear or logarithmic.

Origin:

The reason for the neglect of the equations (3) to (5) is probably that one has a somewhat restrained relationship to the quantities μ and S. One consequence is that many interesting questions that could be answered with their help simply aren't asked in the first place.

Disposal:

Even without or before treating the caloric equation of state, equations (3) to (5) should also be introduced. A side effect is that the importance of the ideal gas law is somewhat adjusted. Just as one can derive equation (3) or (4) from (1), one can also derive equation (1) from (3) or (4).

Friedrich Herrmann

4.46 False friends

Subject:

The concept of heat is introduced and used in textbooks in the following way:

> "If we need hot water, we must supply heat to the water, for example by placing it on a hot plate or heating it with an immersion heater. The water absorbs heat and its thermal energy increases."
>
> "The heat energy *Q* supplied to a body is measured by the change of its internal energy *U*. In the following experiment, heat energy *Q* is generated by friction, which leads to an increase in temperature."

Deficiencies:

I'm not primarily concerned with the concept of heat here. I'm concerned with a behavior of the authors that I find insincere. A little detour first.

Some topics of physics are more difficult, others easier. Something is easy to understand and easy to explain if we have a well-working model – for example, the theory of electricity: "Imagine the electric charge as a stuff that flows through the wire and through the light bulb…"

And there are topics that are more difficult, because one cannot find a suitable model; there is no "it's like something you know". Examples are phenomena from quantum physics, (the wave-particle character of electrons) or relativity (the merging of space and time). If we do not have a suitable model for a phenomenon, we have no choice but to explain to the students: "What I am telling you is strange, almost unbelievable. You have never seen or experienced anything like it, but it is not contradictory! You just have to get used to it. This is the way the world is made."

There is yet a third type of statement: they come across as if they were easily understandable, but they are not, and the difficulty is concealed from the pupils. One deliberately lets them draw a wrong conclusion, but believes that one has kept a clear conscience, because one has not said anything wrong.

Our quotes are an example of this.

- "The water absorbs heat and its thermal energy increases."
- "The heat energy *Q* supplied to a body is measured by the change of its internal energy *U*."

Of course every student understands the sentences in this way: After it has been absorbed by the water the heat is contained in the water, or the heat can somehow be measured, after it has been supplied,.

However, as the authors of the sentences, and hopefully the teachers who teach the subject know, this is not true. And it is not that what has been supplied just changes its name. Rather, it is simply pointless to speak of a heat content.

But the wrong conclusion is inevitable, because it is based on the fact that learners interpret the written or spoken word as it corresponds to the semantics of our language: something released by A was in A before the release and not anymore afterwards, and something received by B was not in B before it is received and it is in B afterwards.

Such statements (and there are several other examples) are one of the reasons why physics (together with chemistry) has become the most hated school subject.

And within physics, thermodynamics performs particularly poorly.

Here is the result of a survey I made with about 20 students studying to become teachers. They were asked how competent they felt in five areas of physics. They rated themselves on a scale from 1 to 5 (1 = very competent, 5 = very incompetent):

mechanics	1,5
electromagnetism	2,7
oscillations and waves	2,8
thermodynamics	4,2
modern physics	3,8

Of course, there are other areas of life in which even the most unlikely things are told in pleasing words, in the hope that the addressee will not check the consistency of the message, but simply repeat it. We think that physics should keep its distance to these areas.

Origin:

The origin in this particular case has been discussed earlier [1-5]. My main concern is our willingness to tell something that we know will be misunderstood. Apparently we all have a tendency to do so.

While studying, we are whipped through physics, and we simply cannot afford to look so closely at every detail that we realize that some problem has been cleverly hidden.

Finally, a suspicion. The experiments, which one makes in school or in the laboratory of the university, I mean the calorimetric ones, in which the specific heat capacity for water and perhaps still other materials is determined, suggest, or even seem to prove, that it is reasonable to conclude that the heat is contained in the bodies: You add *x* kJ of heat to a body, measure the change in temperature, and if you want to return to the previous state, you have no choice but to extract the same *x* kJ from it again in the form of heat. Why should it be wrong to say that *x* kJ of heat is contained in the body and that you can calculate its value from the change in temperature? Everyone who has come to this conclusion must be convinced that the statements in the quoted sentences, namely that it is not the heat but the thermal or internal energy that increases, are only verbal conventions. However, the cause of the fallacy is clear: since liquid and solid substances do not or almost not change their volume when entropy is added, the error in the conclusion cannot be recognized. It could be recognized if looking at gases – but the measurement of their specific heat capacity is not part of the program.

Disposal:

First, two general remarks:

- As a learner: Admit to yourself when you do not understand something.
- As a teacher: Do not hide difficulties behind pleasing words.

More specifically regarding the heat content:

- Explain it correctly. The only way to do that is to talk about gases, but this is probably only recommendable for the university.
- Do not introduce the quantity *Q* at all. The name heat is a false friend. Everything is easier without it. It's like the second black sheep among the physical quantities: work, which, thank goodness, has already been thrown out of some school books and curricula.

[1] F. Herrmann and G. Job, *Historical Burdens on Physics*, Edition 2019, 4.2 State variables

[2] F. Herrmann and G. Job, *Historical Burdens on Physics*, Edition 2019, 4.6 Amount of heat and heat capacity

[3] F. Herrmann and G. Job, *Historical Burdens on Physics*, Edition 2019, 4.8 The equivalence of heat and work

[4] F. Herrmann and G. Job, *Historical Burdens on Physics*, Edition 2019, 4.9 Thermal energy

[5] F. Herrmann and G. Job, *Historical Burdens on Physics*, Edition 2019, 4.10 Internal energy and heat

Friedrich Herrmann

4.47 Free energy

Subject:

> "Free energy, symbol F ..., a thermodynamic state variable that characterizes the ability of a system to perform work. It is defined as $F = U - TS$, where U is the internal energy, T is the temperature in K and S is the entropy."
>
> "The free energy, also Helmholtz potential, Helmholtz free energy or Helmholtz energy after Hermann von Helmholtz, is a thermodynamic potential. It has the dimension of an energy. Free energy is an extensive quantity.…
>
> In thermodynamics, thermodynamic potentials are quantities whose information content completely describes the behavior of a thermodynamic system at equilibrium."

Deficiencies:

1. In which context and for what purpose does one introduce free energy? Our first definition seems to tell it: If we want a gas to perform work, we also want to know how much work it can perform. So we look at its free energy. A first problem about this is that the value of the internal energy contains an arbitrariness, because the zero point can be chosen arbitrarily. Of course, one could simply take the total energy of the gas at rest. But in this case the rest energy is included, so the value is gigantic. It certainly does not answer our simple question. Of course, we did something wrong – the definition was not meant that way. Rather it was meant: The difference of the free energy between two states tells us how much work the gas can perform in a transition between these two states.

But here comes the next problem: Free energy tells us, one might think, how much work is done, for example, by the steam that is expanding in the cylinder of the steam engine of a locomotive. But no, it does not say that. The statement refers only to processes with constant temperature.

However, let us leave our concerns aside for the moment and consider the energy balance in the isothermal expansion of an ideal gas, Fig. 1.

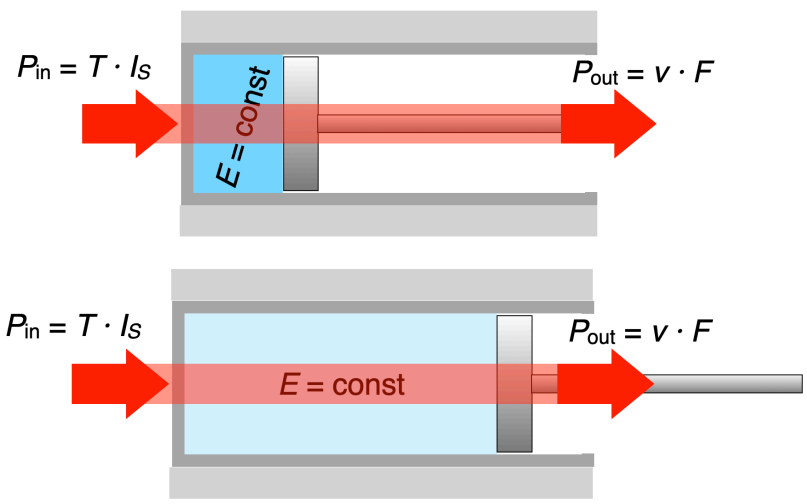


Fig. 1. Expansion of an ideal gas at constant temperature. An energy flow is traversing the gas.

The piston moves to the right, the gas is in thermal equilibrium with the environment, the entropy exchange is done through the left wall of the cylinder.

Since the internal energy of the ideal gas depends only on the temperature, it remains constant during expansion. An energy flow, accompanied by an entropy flow, enters the gas and the same energy flows out of the gas via the piston rod, together with a momentum flow.

This simple process is described with the help of the free energy rather confusingly. The intention is probably that one does not want to say that the outgoing energy (which is called work) entered the gas on the other side. One wants to say that something is coming out, that had previously been within the gas. One would like to attribute the out-flowing energy to a change of some content of the cylinder. But the content of what? The energy in the cylinder does not decrease. So one constructs a quantity, which achieves what is wanted and of which one says, it is an energy. It is not the normal energy, but "free energy".

2. If one is interested in practical applications, yet another question might arise: How can a gas perform work if its pressure is equal to the ambient pressure? The question of how much work a gas (or other system) can perform is easily answered. It is given by the quantity *exergy*, well known to engineers. Physicists do not seem to like the exergy, because its value is not only determined by the state of the working medium, but also by the state of the environment.

3. Just as we learn that free energy is the work content (when running a process at constant temperature), we also learn that enthalpy describes the heat content (when running a process at constant pressure). Doesn't that match up beautifully? Work and heat are the two (only) "process quantities" of physics with its well-known unpleasant properties. They cannot be said to have any value in a given state; one cannot say that a system contains whatsoever amount of work or whatsoever amount of heat. So it seems that one has solved the problem: work input or output manifests itself in a change of free energy, heat input or output in a change of enthalpy. Is it not possible to formulate a balance equation for the work or for the heat in this way? Unfortunately not, because it remains the annoying secondary condition "at constant temperature" or "at constant pressure".

4. With regard to our second quotation: Free energy is a thermodynamic potential. Also here something is not quite coherent. Apart from the fact that one might have the expectation that a potential is a local quantity, i.e. that its value refers to a point in space, the question arises: What is free energy: is it a physical quantity whose value is unambiguously determined by the state of the system under consideration, or is it a function? In the quoted definition the function $F(T,V,n)$ is meant. For if the quantity $U - TS$ is given as a function of, say, S, V and n, it would not be a thermodynamic potential. By the way, the same inaccuracy is found in the analogous mechanical case. The Lagrangian $L(q_i, v_i)$ is a Legendre transform of the Hamiltonian $H(p_i, v_i)$. But one may also find (for instance in Wikipedia) a definition like:
„For systems … the Lagrangian is.

$$L = T - V,$$

where T is the kinetic energy and V the potential energy of the system under consideration."

However, T and V are not functions in the first place. One can write the kinetic energy as a function of the velocities of the particles or as a function of the momentums. The Lagrangian is a function of positions and velocities – otherwise it would not be a Lagrangian.

Moreover, other potentials, for example the electric potential, the magnetic scalar and vector potential or the chemical potential have a well-defined value in a given state. They do not become potentials only when expressed as a function of certain variables.

5. Thermodynamics is considered difficult and it is not very popular. In school, it is part of the curriculum, but it is hardly covered in class, and there are no high school graduation exams on thermodynamics. It also seems to be unpopular at the university. What might be the reason? After all, mathematically it is not very demanding: no nonlinear differential equations, no tensors, not even vectors, no cognitive conflicts as in relativity, no high-dimensional spaces as in quantum mechanics. And one has to admit: It is, if well presented, of a certain elegance.

I think the answer is simple: it is the abstractness of the magnitudes with which it operates: the process quantities heat and work (one could also call them un-quantities), and the quantities H, F, and G, or the functions $F(T,V,n)$, $H(S,p,n)$ and $G(T,p,n)$.

Origin:

The construct of thoughts originates essentially from Helmholtz. Its mathematical aesthetics can be recognized even better in more modern representations [1] .

However, the whole scenario would not have come about:

- if Carnot's caloric, which is essentially identical to our modern concept of entropy, had been retained as a measure of heat instead of defining heat as the energetic differential form δQ ;
- if we had placed the chemical potential (introduced by Gibbs) in the center of our teaching, a descriptive and benign quantity, which can be interpreted as a universal drive for reactions, phase transitions, diffusion and other processes.

Disposal:

One operates with the extensive quantities energy, entropy, amount of substance, momentum, electric charge ... , as well as the corresponding ("energy conjugated") intensive quantities. All these quantities have a descriptive meaning: the extensive ones have the character of amounts, i.e. one may deal with them like with the amount of an imaginary fluid. A local balance can be established for each of them. The intensive quantities have the character of driving measures.

Then, the free energy is no longer needed – as well as the enthalpy H and free enthalpy G.

[1] H. B. Callen, *Thermodynamics*, John Wiley & Sons New York, 1960

Friedrich Herrmann

4.48 Latent and sensible heat

Subject:

"Latent heat" seems to be an important concept in the context of phase transitions. Here some quotes from respected scientific journals:

> "The non-water part of the air carries much less heat than the few percent of water, it seems. One is a transport of sensible heat, the other latent heat."
>
> „The vapor carries with it a form of energy called latent heat. … The energy contained in latent heat is substantial; …"
>
> „Since during freezing, the temperature of the water initially remains unchanged because energy is released as latent heat, …"

Deficiencies

One knows what is meant. But it is difficult to reconcile what is meant with what these sentences actually say. There are two causes for this.

One of them has already been mentioned here several times: In the first and second quotes, it is said that the waterless portion of the air carries less heat. If the air carries heat, then the heat would have to be contained in the air, which, as is well known, is not the case.

But that is not our topic here. Rather, it is about another problem that is manifest in the sentences.

Thus, the first quotation says that the waterless fraction carries much less heat than the few percent (gaseous) water. Both gases are quite ordinary, in good approximation ideal gases. Despite the same temperature, the small part of the gas carries (contains?) more latent heat than the large one.

Something can't be right here. I am in no way accusing the authors of not understanding the processes they are describing. I accuse them of not expressing clearly what they want to say. However, one must admit that it is not easy to express what one wants to say with the help of the unfortunate "process variable" *Q*. Therefore, let us rather analyze the problem with the help of the entropy.

So: water is evaporated at the ground, the water vapor mixes with the air, both together go skyward. Both are gases and both carry entropy – a lot compared to the same amount in the liquid state. However, the entropy contribution of the water is small compared to that of the nitrogen and the oxygen because the percentage of water vapor in the air is small. At high altitudes, the water condenses. In doing so, it gives off more than half of its total entropy; nitrogen and oxygen do not condense, they retain their entropy. The entropy that the water gives away, would be called latent entropy, as long as it was still contained in the water.

Now, in principle, the nitrogen and the oxygen could have been liquefied afterwards (with the help of a refrigerating machine). Then, in a way, a large part of the entropy of the nitrogen and the oxygen would have become latent entropy retroactively.

And finally, one could also argue that the latent entropy of the water was considered too low, because one could have let the water freeze, so that snowflakes or hailstones are formed.

How does one come to call the entropy of the water latent, but not that of the residual air? It is called latent because the water will probably become liquid later. And the entropy of the nitrogen and the oxygen is not called latent, because these gases will probably not liquefy in the near future.

Latent entropy, and even more so latent heat, is therefore not a physical quantity that characterizes a state. Being latent only expresses that something could happen to the gas in the future.

One can try to escape from the logical trap by expressing oneself sufficiently vaguely. But one probably does that anyway. Let's have another look at the second quotation: "The energy contained in latent heat…". Is here one physical quantity contained in the other? Is perhaps even energy contained in the energy?

Origin

The concepts and their names latent heat (also "concealed heat") and sensible heat go back to Joseph Black [1]. Black's heat was a state variable. It is not to be confused with the differential form δQ, to which physicists gave the name heat 50 years later. It rather corresponds to the quantity called entropy 65 years later [2]. One of Black's great merits is that he was the first to distinguish between the intensive quantity temperature and an extensive quantity heat. He also correctly described the heat balance in phase transitions. Only his designations latent and sensible were probably an awkwardness. It is not so that one did not notice anything of the heat or that it was hidden (concealed). One should only have accepted that one recognizes the heat content not only by the temperature, but also by the state of aggregation. At the same temperature, steam contains more heat (in the Black sense) than liquid water.

Disposal

The processes of heating and the transitions between solid, liquid and gaseous are described with the quantities temperature and entropy. The relationship is shown by plotting temperature versus entropy, Fig. 1. Entropy is the independent variable because it is the one that is easily manipulated: We add entropy to ice or liquid water, and see what happens. Two things can happen: the temperature changes and/or the state of aggregation changes.

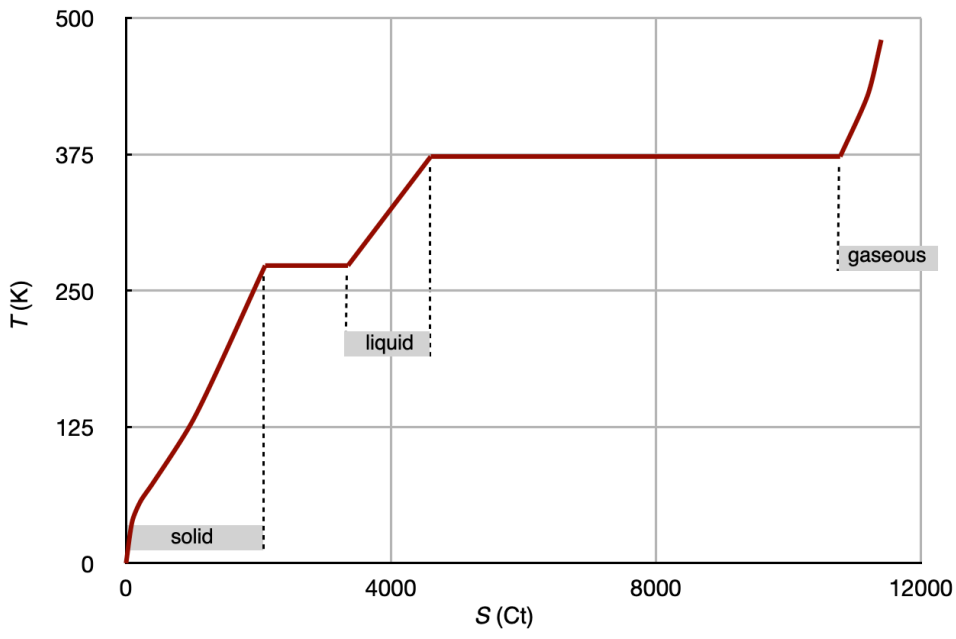


Fig. 1. Temperature as a function of entropy content for 1 kg of water at $p = 1$ bar

[1] J. Black: *Lectures on the Elements of Chemistry,* delivered in the University of Edinburgh by the late Joseph Black, M.D. Published from his manuscripts by John Robison, Edinburgh, Mundell and Son, 1803, S. 120 ff.

[2] G. Falk: *Entropy, a resurrection of caloric a look at the history of thermodynamics,* Eur. J. Phys., 1985, S. 108-115

Friedrich Herrmann

5

Mechanics

5.1 Instantaneous and average velocity

Subject:

In a physics school book I found the following highlighted statements:

"By velocity v of a uniform motion we understand the constant quotient of an arbitrary displacement Δs and the time Δt necessary for this displacement:
$v = \Delta s/\Delta t$."

"For a uniform movement with initial values $t = 0$ and $s = 0$ in addition to $v = \Delta s/\Delta t$ there is: $v = s/t$."

"In reality we find that the instantaneous velocity is approximately equal to the average velocity in a time interval that is as small as possible."

And in another text book, highlighted as well:

"Definition: If for a rectilinear movement of a body the displacement s and the time t are proportional to one another, the constant quotient $s/t = v$ is called the velocity of the body."

"Definition: If in a section of a rectilinear movement all the quotients $\Delta s/\Delta t$ have the same value, then $\Delta s/\Delta t = v$ is the velocity within this section."

"If Δs and Δt are intervals of the displacement and the time of an arbitrary movement, which belong to one another, then

$$\bar{v} = \frac{\Delta s}{\Delta t}$$

is the average velocity of this movement over the displacement Δs or the interval of time Δt respectively."

"The instantaneous velocity at time t_0 is obtained approximately by taking the interval velocity of a time interval which is as small as possible and which contains t_0."

Similar propositions are found for the acceleration.

Such statements are not a peculiarity of those books from which they are taken. They can be found in many other physics text books, whether they are new and have a modern outfit or they are a hundred years old.

Deficiencies:

The concept of velocity is introduced with an unusual meticulousness. Several concerns may arise.

1. The rigorousness cannot be maintained subsequently. It is to compare with the looseness of the introduction of the concepts force, heat or electric current intensity.

2. Just at the beginning of the physics course such a formalization has a daunting effect on the students.

3. The distance is not great between rigorous thinking to pedantry. One may ask if in the present case the limit to pedantry has not been crossed.

4. It is said that velocity is *defined* by $\Delta s/\Delta t$. It is not said that $v = \Delta s/\Delta t$ is the relation between v, s and t. Do we want to say to the students that the concept of velocity which they had before is not the velocity in the sense of physics? We should better not present a mathematically embellished triviality as a new insight. By the way: If one insists to *define* velocity, this can be done yet in another way.[1]

5. What is offered as a way to understand velocity is not really handy. The detour leads over two or three special velocities: the instantaneous, the interval and the average velocity. If with other physical quantities we would proceed in a similar way we would not get very far. Here is what we had to claim when introducing other quantities: "In reality we find that the instantaneous electric current intensity is approximately equal to the average intensity in a time interval that is as small as possible." Or: "The local mass density is approximately equal to the average density in a region of space that is as small as possible."

Origin:

Probably a legacy from the beginnings of physics as a science. In text books from the 18th century one meets a similar meticulousness also in other contexts, where we nowadays do not see any problem.

Disposal:

A disarmament can be reached in several ways. It is not necessary to explain what is meant by velocity and what is meant by constant velocity. The equation $v = s/t$ describes the relation between the velocity, the travelled distance and the time that is needed in the case that the velocity is constant. If it is not constant, we proceed in the same way as we do with other physical quantities whose values changes with time. The velocity is measured with the tachometer.

[1] We sketch two other procedures for defining velocity. However, we do not recommend to use these definitions at school.

1. Velocity can be defined by means of the relation:

$dE = \mathbf{v}d\mathbf{p}$, i.e. energy change per momentum change. This is analogous to the definition of the electric potential difference (energy change per change of electric charge) or that of the absolute temperature (energy change per entropy change).

2. Velocity can be defined operationally. By means of tachometer which must not be calibrated, we can ascertain if a velocity is constant in time. So we can define a unit v_0 of the velocity. Now, multiples of this unit can easily be constructed. A body A is moved with velocity v_0 relative to a body B, which for his part moves with velocity v_0 relative to the earth. Now A has the velocity 2 v_0 relative to the earth.

Friedrich Herrmann

5.2 Acceleration

Subject:

In the context of kinematics it is common to introduce the physical quantity "acceleration". We distinguish between instantaneous and average acceleration, tangential, radial and normal acceleration, angular acceleration, centrifugal, centripetal and Coriolis acceleration. The students also learn that a uniform circular motion is an accelerated motion.

Deficiencies:

1. Technical terms are useful, because with them a scientific statement can be formulated in a succinct manner. Often one can pack in one word what would otherwise require a whole sentence. Thus, a statement can get clearer when employing a well-defined technical term. However, there is an optimum number or concentration of technical terms. When there are too many, it happens that the understandability gets worse again. A statement may get shorter, but since each technical term has to be defined, the whole text may not. Understandability may get worse, since the reader must know the definitions. An example is the proliferation of distinct names for one and the same physical quantity, the acceleration.

2. The motion of a point can be described by various different functions of time. The most commonly used are position $\mathbf{s}(t)$, velocity $\mathbf{v}(t) = d\mathbf{s}/dt$, and acceleration $\mathbf{a}(t) = d^2\mathbf{s}/dt^2$. One can introduce even higher order time derivatives. The third order time derivative of the position is sometimes called jerk. The word clearly expresses the meaning of the concept. However, if the intention is not to run riot at the kinematic playground, we may ask which of these functions are really needed.

Let us consider one of the most important types of motion: a motion with uniform acceleration. A description by means of the function $\mathbf{s}(t)$ is, at least for High school pupils, rather complicated, since **s** is a quadratic function of time. $\mathbf{v}(t)$ is simpler, since it is linear with time. Mathematically even more simple is the acceleration, since $\mathbf{a}(t)$ is constant. But it is the third time derivative which displays the greatest simplicity: it is zero for any t. A lot can be learnt when comparing these four functions. If, however, one is only interested in a succinct description of the motion, one will try to limit the discussion to those functions that give the best intuitive access to the phenomenon. In our opinion these are $\mathbf{s}(t)$ and $\mathbf{v}(t)$. In our example $\mathbf{v}(t)$ tell us, that the velocity increases uniformly with time. We believe that this statement is easier to grasp than that of saying the acceleration is constant. This becomes evident when asking technically versed people to specify the performance of a car. They do not say that the car's maximum acceleration amounts to a certain number of meters per squared seconds, but they say that the car accelerates from 0 to 100 km/h in so many seconds. Thus they argue with velocity instead of acceleration. They express acceleration in terms of velocity.

One might believe that a physics teacher simply cannot do without acceleration, since the quantity appears in Newton's second law. Actually, there is no acceleration in Newton's works. He formulates his second law with the time rate of change of the quantity of motion.

3. The name acceleration for the quantity **a** is the cause of several incongruities. Recently in a paper in a physical magazine I read: "Charged particles emit radiation whenever they are accelerated or decelerated or when they change their direction of motion." There is certainly nothing wrong with this formulation. However, the next sentence says: "Particles that move on a circular trajectory – even when their velocity is constant – are accelerated and emit radiation…". Whereas in the first sentence the author distinguishes between accelerating, decelerating and changing direction, in the next sentence each particle without further ado executes an accelerated motion whenever $\mathbf{a}(t)$ is different from zero.

We know this problem from elsewhere. In the colloquial language we often have different expressions for the positive and the negative values of a physical quantity: acceleration and deceleration, pressure and tension, hotness and coldness… Physics, however needs one single name for a quantity.

Origin:

Contrary to popular belief Newton did not use a quantity "acceleration". According to his formulation of the Second law, the change of the "motion" (*motus*) of a body is proportional to the force that is acting on it. He employed the word *motus* as an abbreviation for *quantitas motus*, the amount of motion, today called momentum. Also Huygens did not use a quantity "acceleration" [1]. In a publication from 1754, Euler used the differential quotient d^2s/dt^2, but he gave it neither a proper name nor a proper symbol [2]. The earliest citation of acceleration as a physical quantity that we have found is in the *Opera omnia* by Johann Bernoulli from 1742 [3]. Apparently it was introduced in the course of the increasing mathematization of mechanics which took place after Newton.

Disposal:

We do not introduce a quantity *acceleration*. As kinematics is concerned we limit ourselves to discuss position and velocity as functions of time. But also in dynamics we do not need acceleration. We formulate Newton's second law as $\mathbf{F} = d\mathbf{p}/dt$.

[1] *E. J. Dijsterhuis*: Die Mechanisierung des Weltbildes.– Springer-Verlag, Berlin, 1956. – S. 528.

[2] *L. Euler*: Vollständige Theorie der Maschinen, die durch Reaktion des Wassers in Bewegung versetzt werden. – Ostwald's Klassiker der Exakten Naturwissenschaften, Nr. 182. – Verlag von Wilhelm Engelmann, Leipzig, 1911.

[3] *S. Sambursky*: Der Weg der Physik. – Artemis Verlag, Zürich, 1975. – S. 428.

Friedrich Herrmann

5.3 Actions at a distance

Subject:

Statements like “the Moon is attracted by the Earth”, “the sun exerts a force on the Earth”, “like poles repel each other, unlike poles attract each other”.

Deficiencies:

These statements suggest that there is an influence or action of one body A on another body B without the participation or mediation of a third system that connects A and B. Since the times of the introduction into physics of modern electrodynamics by Faraday and Maxwell, i.e. the first field theory, scientists are convinced that such actions do not exist and that such a description is inappropriate.

Origin:

The action-at-a-distance language that can be found in all physics text books dates from the times of Newton. Indeed, before the theory of Faraday and Maxwell came into being there was no other choice than imagine the electric, magnetic and gravitational forces as actions at a distance. Newton himself considered the actions at a distance a flaw of his theory.

Disposal:

As soon as gravitational, electric and magnetic forces between two bodies are discussed, the corresponding field is introduced as a third participant. The field is described as a system that is as real as the two bodies. The electric attraction or repulsion for instance is described in the following way: Two bodies with like charges are pulled away from each other by the field, bodies with like charges are pulled together.

Friedrich Herrmann

5.4 Newton's laws

Subject:

1. Every body persists in its state of being at rest or of uniform motion in a straight line unless compelled to change its state by the action of an external force.

2. The change of the momentum of a body is parallel and proportional to the force acting on it.

3. The forces of two bodies on each other are always equal and are directed in opposite directions.

Deficiencies:

All of the three laws are special cases of a statement, that can be formulated in a much simpler way: Momentum cannot be created and cannot be destroyed. This is easily seen when taking into account that the quantity "force" is nothing else than the current intensity of a momentum current. Thus, Newton's laws can also be formulated in the following way:

1. The momentum of a body does not change as long as no momentum enters or leaves the body.

2. The time rate of change of the momentum of a body is equal to the momentum current flowing into or out of the body.

3. When a momentum current is flowing from a body A to a body B, the momentum current leaving A is equal to the momentum current entering B.

These corollaries of the law of momentum conservation are so simple that one would hardly attribute to them the status of theorems or laws in their own right. To convince oneself one just has to formulate the corresponding statements for another conserved quantity, or even more simply for an amount of water: "The amount of water in a container does not change as long as no water enters or leaves the container.…"

Origin:

Everybody knows the origin of Newton's laws. However, it needs a thorough analysis of Newton's work to understand why in the Newtonian system the three laws appeared as independent from each other. They are components of a complicated system of observations and definitions. Of course, Newton did not place momentum conservation at the beginning of his reflections.

Disposal:

Introduce momentum at the very beginning of the mechanics classes as a quantity in its own right, as a measure of the "amount of motion", or in more colloquial terms, "drive" or "impetus". When the momentum of a body changes, do not say, "a force is acting on it", but "momentum is flowing into it (or out of it)". This way of speaking is unusual for an experienced physics teacher. For the beginner, however, it is easier, since it avoids some of the complications that the discussion of Newton's laws, especially the third law brings with it.

Friedrich Herrmann

5.5 Static equilibrium and Newton's third law

Subject:

Forces act on bodies. If a body P, on which another body A exerts a force F_{AP}, is not accelerated, then there must exist at least one other body B which exerts a force F_{BP} on P in such a way that the resulting force on P is zero. P is in a state of static equilibrium. Now, when A exerts a force F_{AP} on P, according to Newton's third law P must exert a force F_{PA} on A. Correspondingly, P must also exert a force F_{PB} on B. All of the four forces F_{AP}, F_{BP}, F_{PA},and F_{PB}, have the same absolute value, whereas the directions are pairwise opposite:

$F_{AP} = -F_{BP}$, $F_{AP} = -F_{PA}$, $F_{BP} = -F_{PB}$, $F_{PB} = -F_{PA}$.

This is the description of the situation in which P is at rest. It is, a part from the case that all of the forces are equal to zero, the simplest static situation that we can imagine. Who wants to understand what a force is, must be able to conceptually distinguish between these four forces.

Deficiencies:

The treatment of the problem is so complicated, that an average pupil will hardly understand it. Actually, bachelor students have problems with distinguishing these forces. Even so, school makes the desperate effort to introduce Newton's concept of force in the lower secondary school.

Origin:

The concept of force that we employ still today was conceived by Newton. It was introduced in an epoch when mechanical interactions could only be described by actions at a distance. The concept of field was introduced in physics only more than a hundred years later. According to Newton a force is attributed to two bodies: the body that exerts the force and the body on which the force is exerted. In the example that we have mentioned, there are three bodies, which leads to six forces. Four of them have to do with Body P, i.e. they are either exerted by body P or body P is exerting them.

A conceptual simplification could have been introduced in the middle of the 19th century when Faraday and Maxwell introduced the field concept. Actually this was done only after it had become clear that momentum should be considered as a quantity on its own right, instead of an abbreviation for the product $m \cdot v$. Indeed, in 1908, i.e. three years after the publication of the Special Theory of Relativity, Max Planck [1] has shown that a force is nothing else than the intensity of a momentum current. Thus, the value of a force would not refer to two bodies, but to a sectional area of the momentum conducting system (in the same way as other currents refer to cross-sectional areas).

If this insight is applied to the situation cited above, the huddle of forces simply goes away. All of the four forces turn out to be the momentum current intensity of the same current considered at four different cross sections: F_{PA} is the intensity of the current that flows from A and P when leaving A, $-F_{AP}$ is the intensity of the same current when entering P, F_{BP} is the intensity of this current when leaving P again, and $-F_{PB}$ finally is its intensity when entering B. Since nowhere, neither within the bodies nor between them, momentum does accumulate, the absolute value of all of these intensities is the same. The algebraic sign is not always the same because the surface area to which the current refers is not alway oriented in the same direction.

Disposal:

The whole spook disappears when, instead of forces, momentum currents are used. Then the verbal description of the situation is as follows: A momentum current flows from A to P and from P to B. Since no momentum accumulates anywhere, the current intensity must be the same at every cross section through the current.

[1] *M. Planck*: Phys. Z. 9, 1908, p. 828.

Friedrich Herrmann

5.6 Absolute space

Subject:

“Absolute space, in its own nature, without regard to anything external, remains always similar and immovable.” (Newton [1])

“It is not necessary to mention, that Newton with the reflections that have just been reported acts against his intention to study only the objectively existing entities. Nobody can say anything about the absolute space and the absolute movement; they are no more than mental constructs, which have no correspondence in our experience.” (Mach [2])

“Therefore it was necessary to debunk the words “absolute time” and “absolute space” as unproved, magical prejudices of prerelativistic times.” [3]

Deficiencies:

Gravitational phenomena can be divided into two classes.

Those of the first class have to do with the gravitational interaction of bodies that are at rest relative to each other. In classical physics they are described by Newton’s law of gravitation and they are similar to the phenomena of electrostatics. We shall call them gravitatostatic phenomena. A body feels that component of the gravitational field that is described by the vector field of the gravitational field strength.

The phenomena of the second class are observed when a body is accelerated. They are described today (together with the statical phenomena) by the metric tensor of the General Theory of Relativity. A body “feels” when it is accelerated relative to other bodies. The distance dependance is different from that of the statical interaction: bodies at great distances have a greater weight than in the case of the statical forces [4]. For that reason acceleration forces appear only relative to the huge masses that are distributed in the universe, whereas the influence of the “small” mass in our neighborhood is so minuscule that it could not yet be detected. They would show up in the Lense-Thirring-Effekt (also called gravitomagnetic effect). Physicists are convinced that the effect exists.

How did Newton deal with the two classes of forces or effects? He described the first one, the gravitaostatic effects with his well-known law of gravitation. He considered it a deficiency that his theory suggested an action-at-a-distance view, and he uttered his view unmistakably. However, his uneasiness did not let him go as far as to mention in this *Principia* a medium that could be made responsible for the transmission of the momentum (the *quantitas motus*) between the celestial bodies. We all know his “*hypotheses non fingo*”. One might believe that from his point of view it would have been consequent to see the cause of the second class of forces, the inertial forces, in the stars, as it was indeed proposed by the somewhat younger George Berkeley. However, here Newton preferred another interpretation, which actually is the more sound idea. The acceleration forces do not originate in remote bodies, but in the “absolute space”, i.e. in something that exists at the same place as the considered body. He thus uses a local description. With a certain right we can consider this idea a precursor of what later was called a field.

From this point of view one would say that Mach’s critique is not adequate, and that an unreflected condemning of the absolute space –see our third citation– does not help for a detached assessment of Newton’s ideas.

Origin:

It was already mentioned in the previous section. One of the reasons that the statical and the dynamical gravitational phenomena are treated in such a different way, may be that an action-at-a-distance theory for the inertial forces was not yet ready, although the basic idea from Berkeley existed already.

Disposal:

A little more respect for Newton’s absolute space. The idea is not so bad as many make us believe.

[1] Wikipedia, keyword “Absolute time and space”

[2] *E. Mach*: Die Mechanik in ihrer Entwicklung. Leipzig: Brockhaus, 1897, S. 223.

[3] Dorn-Bader, Physik, Gymnasium Gesamtband, Hannover: Schroedel, 2000, p. 405.

[4] *D. W. Sciama*: The Physical Foundations of General Relativity. New York: Doubleday & Company, 1969, p. 22-33.

Friedrich Herrmann

5.7 Momentum as the product of *m* and *v*

Subject:

Usually the momentum of a body is defined as the product of its mass and its velocity:

$$\boldsymbol{p} = m \cdot \boldsymbol{v} \,. \qquad (1)$$

Thus, $\boldsymbol{p}$ is nothing else than an abbreviation for the product of *m* and *v*. Momentum seems to be a typical example of a "derived quantity". In some books momentum is explicitly called an auxiliary quantity [1].

Deficiencies:

There are several arguments to introduce momentum not as a derived but as a basic quantity in its own right.

1) Momentum is a conserved quantity. This property makes it easy to measure the momentum of a moving body without recourse to equation (1) [2,3]. Since (gravitational) mass and velocity can be measured independently, equation (1) can be verified experimentally.

2) Equation (1) does not hold for every system. The momentum of the electromagnetic field cannot be calculated with this equation. The momentum density of the electromagnetic field can be calculated form the electric and the magnetic field strength:

$$\rho_p = \frac{\boldsymbol{E} \times \boldsymbol{H}}{c^2}$$

3) There is a far reaching analogy between mechanics and electricity: a correspondence between physical quantities and relations between these quantities. For example, the electric analog of the conserved extensive quantity momentum is electric charge, and the analog of the intensive quantity velocity is the electric potential. The analog of equation (1), which tells us that for non-relativistic velocities momentum is proportional to the velocity, is the equation

$$Q = C \cdot U, \qquad (2)$$

which tells us, that for a capacitor with fixed plates, the charge is proportional to the potential difference between the plates. A comparison of equations (1) and (2) shows that mass can be interpreted as "momentum capacitance". The greater the mass of a body is, the more momentum it contains at a given velocity.

The comparison shows that it is no convenient to define momentum by equation (1). This is as if one would define electric charge by equation (2), instead of introducing it as a quantity in its own right, which can be measured without recourse to *U* and *C*.

4) To introduce momentum directly as a self-contained quantity is also suggested by the fact, that momentum (or more exactly momentum density) is a component of the energy-momentum tensor. That means that for the gravitational field momentum plays a similar role as electric charge for the electromagnetic field. Together with the energy density, the energy flow density and the momentum flow density, it belongs to the sources of the gravitational field. The sources of a field play an important part in the fundamental interactions, and it seems not convenient to consider them as derived quantities.

Origin:

In contrast to the electric charge, the physical quantity momentum came into existence in a long historical process. In the 17th century it was a professed aim of the mechanical sciences to formulate the laws that govern collision processes. It was correctly expected that an invariant quantity should play a decisive role and it was tried to express this quantity as a combination of mass and velocity.

In 1644 Descartes published his Principia philosophiae, in which he claimed the conservation of the product of mass and velocity, the *quantitas motus*, the amount of motion. Some decades later Leibniz believed to prove that the product of the mass and the square of the velocity is the "correct" invariant in a collision process. As a consequence the famous, long-lasting dispute about which is the true "measure of force" broke out, which was brought to an end only in 1726 by Daniel Bernoulli, and in which there were no winners and no losers. What happened was the emergence of two quantities, one of which is what we now call momentum and the other kinetic energy.

Naturally, the result was a momentum that was *defined* by equation (1). Only much later it was discovered that if a conserved quantity momentum is to be constructed relation (1) has to be abandoned. The Theory of special relativity tell us that the new, conserved momentum is not proportional to velocity. Equation (1) was saved by introducing a velocity-dependent mass.

Disposal:

Introduce momentum as a quantity in its own right, with its own measuring procedure, i.e. in the same way as we are used to introduce electric charge. Then equation (1) takes over the role of a definition of the inertial mass, as the factor of proportionality between momentum and velocity.

[1] *R. W. Pohl*: *Mechanik, Akustik und Wärmelehre*.– Springer-Verlag, Berlin, 1969.– S. 45

[2] *F. Herrmann*: The Karlsruhe Physics Course, The Teacher's Manual, p. 23,
http://www.physikdidaktik.uni-karlsruhe.de/kpk/english/KPK_Teacher.pdf

[3] *F. Herrmann, M. Schubart*: Measuring momentum without the use of *p* = *mv* in a demonstration experiment, Am. J. Phyis. **57** (1989), p. 858

Friedrich Herrmann

5.8 Momentum underrated

Subject:

"Second law: The acceleration $\boldsymbol{a}$ of a body is directly proportional to the force $\boldsymbol{F}$ acting on the body and inversely proportional to the mass m of the body, i.e.;

$\boldsymbol{F} = m \cdot \boldsymbol{a}$."

"The magnitude of the centripetal force on an object of mass m moving at tangential speed v along a path with radius of curvature r is:

$F = ma_c = \frac{mv^2}{r}$."

Deficiencies:

Clearly there is no objection to the validity and to the usefulness of the equation

$$\boldsymbol{F} = m \cdot \boldsymbol{a} \,. \qquad (1)$$

We believe, however, that it is not convenient to call it "Second law", or "Fundamental law of motion", since it subsumes two other law that should be clearly kept apart from one another. The first one is indeed Newton's second law

$$\boldsymbol{F} = \frac{\mathrm{d}\boldsymbol{p}}{\mathrm{d}t} \,. \qquad (2)$$

It tells us that momentum can change only when a momentum current is entering or leaving a system (in other words: when "a force is acting on the system"). It thus claims that momentum is a conserved quantity.

In order to obtain equation (1) we still need the equation

$$\boldsymbol{p} = m \cdot \boldsymbol{v} \,, \qquad (3)$$

or its time derivative

$$\frac{\mathrm{d}\boldsymbol{p}}{\mathrm{d}t} = m\frac{\mathrm{d}\boldsymbol{v}}{\mathrm{d}t} \,.$$

The character of equation (3) is different from that of equation (2). It is what in another context one would call a constitutive equation. Such equations are valid for certain systems under certain circumstances. So, the equation

$\boldsymbol{p} = m \cdot \boldsymbol{v}$

is valid only as long as the velocity is small when compared with c, and it is not valid for the electromagnetic field, since this system is not described by the variables m and $\boldsymbol{v}$.

In order to make our argument clearer consider the corresponding electric laws. The following analogy between laws of mechanics and electricity is well-known:

mechanics	*electricity*
$\boldsymbol{F} = \mathrm{d}\boldsymbol{p}/\mathrm{d}t$	$I = \mathrm{d}Q/\mathrm{d}t$
$\boldsymbol{p} = m \cdot \boldsymbol{v}$	$Q = C \cdot U$
$\boldsymbol{F} = m \cdot \mathrm{d}\boldsymbol{v}/\mathrm{d}t$	$I = C \cdot \mathrm{d}U/\mathrm{d}t$

It is obvious, that to the equation

$I = C \cdot \mathrm{d}U/\mathrm{d}t$

one would not give a name like "fundamental law of electricity".

When skipping equation (2) and declaring that equation (1) is the fundamental law, or also the Second law, then there is no need to mention momentum. Apparently, this is considered an advantage. Indeed, in the usual course of teaching mechanics the quantity force, i.e. momentum current, is discussed extensively at the beginning, whereas momentum has to wait until collision processes are discussed.

The tendency to circumvent momentum can also be observed at other instances, see our second citation. Here again the intermediate result is missing. The first step to get the centripetal force is to calculate the time rate of change of momentum of the rotating body:

$$\frac{\mathrm{d}p}{\mathrm{d}t} = m\,\frac{v^2}{r} \,.$$

The term on the right hand side of the equation only contains quantities that characterize the rotating body. Only by using Newton's second law we get:

$$F_c = \frac{mv^2}{r} \,.$$

Origin:

The disregard of momentum and its reduction to a mere invariant in collision processes is not a relict of the early times of mechanics. On the contrary, already before Newton's time and at Newton's time momentum was that quantity in whose balances one was interested. Its original latin name was *quantitas motus*, which can be translated as *quantity of motion* or also as *amount of motion*. Momentum gained importance in modern physics. Relativistic physics tells us that momentum density is a component of the energy-momentum tensor, and thus belongs to the sources of the gravitational field. Therefore, the apparent disregard of momentum in elementary mechanics seems incomprehensible.

We believe that we have to blame Leibniz. In the famous controversy about the "true measure of force" between Leibniz and the Cartesians the question was which of the two expressions $m \cdot v$ and $m \cdot v^2$ is the "correct" measure of the amount of motion. Today we know that physics needs both expressions. The first one is what we call today momentum and the second (apart from a factor of 2) is the kinetic energy. Both of them correspond to what at Leibniz's time was called force by some, and what today we would call impetus or drive or momentum (in the colloquial sense of the term). Now, in the teaching tradition of mechanics, this everyday concept of impetus was primarily associated with the physical quantity kinetic energy, i.e. essentially with Leibniz's $m \cdot v^2$. The simple reason is that we generally introduce kinetic energy before momentum. Thus, when momentum is introduced the place for a physical quantity that measures what we intuitively associate with the concept of impetus is already occupied. As a result, most students consider momentum as the more abstract quantity. They learn that momentum measures something rather similar to kinetic energy without being the same quantity. Thus momentum is perceived as a more difficult quantity and its role is essentially reduced to an invariant in collision processes.

By the way: In this respect the fate of momentum is somewhat similar to that of entropy. Initially this quantity was a perfect measure of what in colloquial term would be called heat. After the introduction of energy and the discovery of its conservation the name and the mental picture that it associated with the word heat was transferred to the differential form $\mathrm{d}Q$. As a result entropy was now considered a rather "abstract" quantity.

Disposal:

Introduce momentum as a basic quantity right at the beginning of mechanics. Introduce it as a measure of what in colloquial terms would be called impetus.

Introduce Newton's second law as Newton did it: $\boldsymbol{F} = \mathrm{d}\boldsymbol{p}/\mathrm{d}t$. Introduce the relation $\boldsymbol{p} = m \cdot \boldsymbol{v}$ later, in the same way as you discuss the relation $Q = C \cdot U$ after the introduction of the electric charge Q.

Regarding the circular motion, before introducing the centripetal force show that

$$\frac{\mathrm{d}p}{\mathrm{d}t} = m\,\frac{v^2}{r} \,.$$

Friedrich Herrmann

5.9 Impulse

Subject:

"The impulse $\Delta\vec{p}$ of a force is a vector quantity that is defined by

$$\Delta\vec{p} = \int_{t_0}^{t_1} \vec{F}\, dt \quad ."$$

Deficiencies:

The introduction of the impulse begins with the equation $\vec{F} = d\vec{p}/dt$. Each of the three quantities $\vec{F}$, $\vec{p}$ and t has a clear meaning. The equation is transformed into $d\vec{p} = \vec{F}dt$, and a name is given to the expression on the right hand side. In text books this is often done in all detailedness, and it is suggested that there is something to understand that exceeds Newton's second law. In particular, the student gets the impression that a new physical quantity has been introduced. Actually, the impulse is not a physical quantity – at least not in the usual sense of the word. A physical quantity has in a given state a well-defined value [1]. This, however, is not true for the impulse.

That the concept is not indispensable can also be seen in the fact that various other "quantities" could be constructed by using the same recipe – what is not done. Indeed, when considering that each force can be considered a momentum current and impulse the time integral of the momentum current, then it is seen that corresponding integrals can be written for any other current, for instance an electric current, a mass current or an energy current. So, the expression

$$\int P dt$$

(where P is the energy current or "power") could be introduced as a new "quantity" and a proper name could be given to it.

We do not want to say that it does not make sense to calculate such an integral. We only believe that it is not convenient to present it as a new physical quantity.

Actually, it is trivial that when integrating a force (a momentum current) with respect to time the momentum change of the body, on which the force is acting (to which the momentum is flowing) will result. It is trivial because momentum is an extensive or substance-like quantity.

When multiplying the current strength of the water current as one is filling the bathtub with the time, one gets the amount of water that has flown into the tub. In order to understand this statement we do not need the concept of a "water impulse".

A proper name for the time integral of the force, is as superfluous as the names work and heat for the differential forms $\boldsymbol{F}d\boldsymbol{s}$ and TdS respectively. These too are expressions that do not represent physical quantities in the usual sense.

Origin:

The expression was introduced to describe momentum transfer processes that are short in time and whose particular time dependence was not important. This concern is understandable, but it can also be met without a new name. It is sufficient to say that a certain amount of momentum is transferred. However, among the quantities force and momentum, force had always been considered the more fundamental one. Momentum was only conceived as an abbreviation for the product of mass and velocity. So it seemed more natural to make a statement about the force instead of momentum.

Disposal:

When taking momentum as a quantity of its own right seriously, impulse is not needed. Mechanics does not lose anything when omitting the concept, but it gains clarity.

[1] *G. Falk*: Theoretische Physik, II Thermodynamik, Heidelberger Taschenbücher. Springer-Verlag Berlin, 1968, p. 4

Friedrich Herrmann

5.10 State of motion

Subject:

Students know what to answer when asked in an exam which are the effects of a force: A force causes a deformation or a change in the state of motion of the body on which it acts.

Deficiencies:

By “change of the state of motion” is meant a change of the velocity. Secondary High school students know vectors. So we can expect that they know that velocity is a vector quantity. When saying that the state of motion is changed we express something vaguely what could have been said clearly: by specifying the physical quantity whose value is changing, as is common practice in other circumstances. We say that when supplying heat the temperature increases. We do not say that the thermal state changes. When supplying electric charge to a body we say that the electric potential increases. We do not say that the electric state changes. When inflating a tire we say we supply air to the tire. We do not say that we change the compressional state of the tire.

Origin:

The wording comes directly from the great master. His first law reads:

“Corpus omne perseverare in statu suo quiescendi vel movendi uniformiter in directum, nisi quatenus a viribus impressis cogitur statum illum mutare.”

(Every body persists in its state of being at rest or of moving uniformly straight forward, except insofar as it is compelled to change its state by force impressed.)

Here is the “state of moving”.

It is understandable that Newton had to express himself in this way. He could not refer the velocity as a vector quantity, since the concept vector was introduced only more than a hundred years later.

Disposal:

The expression “state of motion” is suitable in a more general sense. The meaning of “state of motion” might include all the data that characterize its motion: velocity, acceleration, rotation…. However, in the context of Newton’s laws it is better to specify that the change is that of the velocity.

Friedrich Herrmann

5.11 Muscular force

Subject:

"Although we may have already an intuitive idea of a force as a push or a pull, like that exerted by our muscles… Newton's laws allow us to refine our understanding of forces." [1]

"We all have a fairly deep intuitive understanding of what forces are and what effect they have on objects. We are constantly using our muscles to exert forces: we pull up on a coffee cup to get it to our mouths, we push against a car stuck in a ditch in order to get it moving and we exert a force to stop a basketball as we catch it." [2]

"The concept of force can be traced back to our muscular sensation." [3]

Deficiencies:

We perceive the actuation of our muscles as an effort. We do not perceive it as a specific sensory perception but as an act of volition. But for which physical quantity can this effort be considered a measure? On the one hand a *force* is acting as long as the muscle is tended. On the other the muscular activity needs energy. ATP is transformed into ADP, regardless of whether our muscles move something (deliver mechanical work) or not (i.e. only produce heat). Thus, our muscular sensation points to a force (a momentum current) just as much as to an energy current, i.e. to the physical quantity *power*. Since in the teaching of physics the concepts force and power are often confounded, we believe that it is not wise to appeal to the muscular sensation when introducing the concept of force.

Origin:

There is reason to suspect that the muscular sensation is put forward because one might take the muscles for the cause or the causer of the force. Let us consider an example: Who or what is the cause of the force in the string in Fig. 1a? Our immediate feeling may be that it should be the spring. And in Fig. 1b? Shouldn't is be the manikin with its muscles? This feeling brings us to say: "The spring pulls", or "The manikin pulls". We do not say: "The string pulls", or "The wall pulls". But something must be wrong with these statements. We consider Fig. 1c. Here, which of the two springs would be responsible for the force, which one would be the cause of the force? And which one of the two manikins in Fig. 1d? Finally, we can also consider the string to be a spring with a very great spring constant. Thus, our procedure to find the cause of the force does not work. So the question is: But can it be that our feeling has cheated us? Isn't there really nothing that distinguishes the spring or the manikin from the rest of the arrangement? Yes, there is. Both, the spring and the manikin can act as a source of mechanical energy. And therefore they appear to our feeling as the origin, the responsible, the causer of what we observe. Actually, they must not deliver mechanical energy, they only must be able to deliver it. Thus, if we make them responsible for the force, we miss the target.

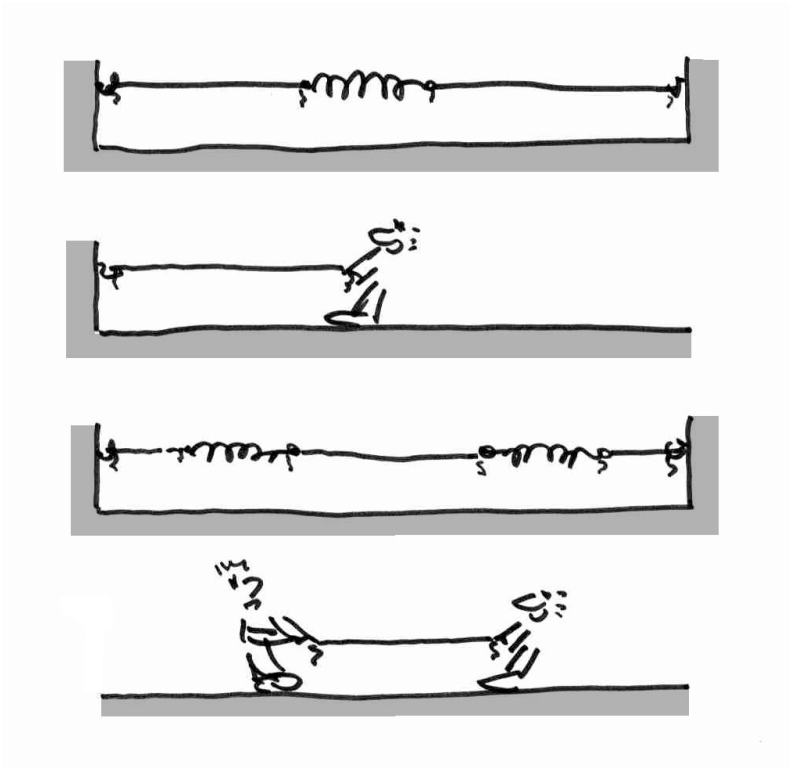


Fig. 1. (a) Is the spring the cause of the force? (b) Is the manikin the cause of the force? (c) Which of the two springs would be the cause of the force? (d) Which of the two manikins would be the cause of the force?

Disposal:

Actually, we are perfectly able to perceive forces that act on our body. Nature has equipped us with special sensory organs for the purpose. In our skin we have sensors for compressive, tensional and shear stress, which are as reliable for the measurement of forces as other sensors in our body are reliable for the „measurement" of other physical quantities: temperature, light intensity and sound intensity. In this way we can feel forces, regardless of whether our muscles are active or not. We can even get a rather good idea of the unit of force: A slight pressure on our skin with one finger, or also a 100-g-weight placed on our arm corresponds to one Newton.

[1] *P. A. Tipler and G. Mosca*: Physics – for Scientists and Engineers, Sixth edition, W. H. Freeman and Company, New York, 2008, p. 93.

[2] http://theory.uwinnipeg.ca/mod_tech/node20.html

[3] Cited from a German school book

Friedrich Herrmann

5.12 Restoring force

Subject:

In school books one finds definitions of the harmonic oscillator, often highlighted, like the following: “In mechanics, a harmonic oscillator is a system that experiences a restoring force which is proportional to the displacement: $F = -D \cdot s$.” (1)

Apparently much importance is attached to this proportionality.

Deficiencies:

1. A highlighted sentence calls our special intention: “What I say is important!” The above-cited statement however, does not merit this status of importance.

We can note, that there are two other variables that are proportional to each other, momentum and velocity:

$p = m \cdot v$ (2)

Both relations (1) and (2) have the same structure and they are of similar importance for the oscillation. They characterize the two components of the oscillator: the spring (equation (1)) and the moving body (equation (2)). Together with the law of momentum conservation they allow to write down the equation of motion.

One may argue that equation (2) is not worth a key sentence since it is true anyway. Deviations from the proportionality between velocity and momentum are observed only at relativistic speeds. Quoting it as a condition for a harmonic movement would appear pedantic, would it? But how about equation (1)? When introducing the oscillator as a system consisting of a massive body and a spring, this equation too is understood, since nobody will intend to overstretch the spring.

In order to look at a mechanical problem from a certain distance, it is good practice to translate it into an electrical problem.

Also an electric circuit oscillates harmonically only if two linear relations hold, which characterize the two components of the circuit, i.e. the capacitor and the coil:

$n\Phi = L \cdot I$

and

$Q = C \cdot U$.

Together with the law of charge conservation they allow to write down the differential equation of the oscillator.

In this case it is particularly easy to violate the linearity: When using a coil with an iron core that goes into saturation or when using an electrolytic capacitor. Despite this possibility nobody believes it is necessary to formulate:

“In electricity, a harmonic oscillating circuit is a system consisting of a coil and a capacitor, where the magnetic flux in the coil is proportional to the electric current and the voltage of the capacitor is proportional to its electric charge”. Why don’t we formulate such a definition? Because we take it for granted that the magnetic flux is proportional to the electric current and the voltage is proportional to the charge.

2. Do we really need a proper name for this force (*restoring force*) in the mechanical oscillator? To be consequent, one should then argue, that the electric current in the coil of the oscillating circuit merits its own name, *restoring current* for instance.

Origin:

This is one more example for the special treatment that receives mechanics. The historically caused dominance of mechanics remains undisputed.

Disposal:

Do not feign rigor where is no need for it. Highlighting sentences can be helpful. Highlighting trivial statements is annoying.

Friedrich Herrmann

5.13 Line of action

Subject:

When introducing the physical quantity “force” it is sometimes stressed that a force is determined by three elements, or that three indications must be given: 1. Its magnitude, 2. its sense, 3. its point of application. Later, in the context of the treatment of the torque a forth characteristic is attributed to the force: its line of action.

Deficiencies:

This is only one example of how the concept of force is presented in an unnecessarily complicated way. Again, force seems to require a special treatment. Force is a vector quantity and as such a force is determined by the three components or alternatively by its magnitude and direction. But why should it be necessary to attribute in addition a point of attack to a force? Is that a particular characteristic of the force? It is not. It is rather a kind of quirk.

The values of most physical quantities refer to one of the following four geometric entities: a point (as pressure and temperature), a line (as voltage), a surface area (as electric current intensity and magnetic flux) or a region of space (as mass, electric charge and entropy). Those quantities that refer to a point are sometimes called intensive or local quantities. Those quantities that refer to a space region are the extensive quantities. The quantities whose values belong to a surface area are the so-called currents, fluxes or flows. This classification holds for scalar, as well as for vector and tensor quantities. So, temperature is a scalar, and electric field strength a vectorial local quantity. Electric charge is a scalar and momentum a vectorial extensive quantity. Power is a scalar, force a vector “surface area quantity”.

We now can say more clearly what’s up with the point of attack of a force. It is supposed to be the geometrical entity to which the value of a force refers. Now, two remarks are indicated:

1. This geometrical entity to which the value of a force refers is not a point but a surface area.

2. It is not common in physics to mention this entity in the definition of a quantity. We do not say: a temperature is determined by its value and the point at which the temperature is considered. And we do not say: the electric charge is determined by its magnitude and the body where it is sitting.

Origin:

In classical point mechanics, as developed in the 18th and 19th centuries, a force corresponds indeed to a point. Point mechanics was and still is very successful, but in particular as far as school physics is concerned, we should be aware that it is a particular approximation in which several concepts that we need in school physics become singular or senseless, as for example all densities and current densities.

Disposal:

It should become clear that the value of a force refers to a surface area. But this must not be stressed as a particular feature of forces. It is equally true for any other current or flow quantity.

Friedrich Herrmann

5.14 Pressure and force

Subject:

"The pressure in a fluid acts with equal magnitude upward, downward and sideways."

Deficiencies:

The citation is taken from a somewhat older university textbook. It illustrates a problem that we can encounter also in more modern texts: the distinction between force and pressure.

Traditionally force is introduced first. Thereafter pressure is defined as force per unit area. However, the relation cannot be

$p = \Delta F/\Delta A$,

since pressure p is a scalar and ΔF is a vector. Moreover, ΔA is a vector, and one cannot divide by a vector.

This problem can be eluded by defining pressure by means of

$p = dE/dV$,

i.e. pressure is energy change per volume change. But normally this is not done. If sticking on the force per unit area introduction, instead of

$p = \Delta F/\Delta A$

one should write:

$$\mathbf{F} = \boldsymbol{\sigma} \cdot \mathbf{A} \,, \qquad (1)$$

where $\boldsymbol{\sigma}$ is the tensor of mechanical stress. In static fluids (without internal friction) this tensor has only three diagonal elements which all have the same magnitude. It can thus be characterized by a single number, which is called hydrostatic pressure. In this case, equation (1) can be written:

$\mathbf{F} = p \cdot \mathbf{A}$,

In general, $\boldsymbol{\sigma}$ does not have such a simple structure. This is obvious, even to the layman: It is possible to expose an object in three mutually orthogonal directions independently to mechanical stress. If this simple observation is not discussed, it will be difficult to understand why it is worth mentioning that the pressure in fluids is the same in all directions.

There is yet another problem with our citation: Apparently the author himself got entangled in the jungle of the concepts scalar, vector and tensor. Indeed, the statement that the pressure is the same upwards and downwards does not make sense. A tensor distinguishes three mutually orthogonal axes, but these axes have no orientation. Therefore, there is no pressure upwards or downwards. There is only a pressure in the vertical direction. Thus vertical stress (or pressure) can be different from horizontal stress (which is not the case in fluids however). Apparently, pressure upwards and downwards has been confused with force upwards and downwards.

Origin:

The old Newtonian idea according to which a force acts on a body. When defining pressure by means of the force, it is rather natural to look for a body on which a pressure is acting, and it seems plausible that pressure (or mechanical stress) has an orientation.

Disposal:

Introduce force as a quantity that refers to a surface area, i.e. not to a point and not to a body.

Before discussing the pressure in fluids introduce mechanical stress in solid materials and show that it depends on the direction. It is easy to see that one can impose mechanical stress on a body independently in three mutually orthogonal directions. Liquids and gases are special cases, in which the three stresses have identical values.

(Another special case are electric and magnetic fields. In this case the three principle stresses have the same magnitude. In the direction of the field strength vector the stress is negative (tension) and in the directions perpendicular to the field strength it is positive.)

Avoid saying "pressure on…". One can say: pressure in a given direction, for instance pressure in the horizontal direction.

Friedrich Herrmann

5.15 Dynamic pressure

Subject:

The Bernoulli equation

$$p + \rho \cdot g \cdot h + \frac{\rho}{2} \cdot v^2 = \text{const}$$

holds for a stationary, incompressible, inviscid (frictionless) fluid. p is the pressure, ρ the mass density, g the gravitational field strength, h the height (positive direction upwards) and v the velocity. Here, “const” means that the sum at the left hand side of the equation does not change when one is moving along a streamline. If the values of the local variables are the same in every point of a cross section, then the condition is even less restrictive. Then “const” means “has the same value at every cross sectional area“.

Usually, the equation is interpreted in the following way. There exist several types of pressures: the static pressure p, the gravitational pressure $\rho \cdot g \cdot h$ and the dynamic pressure or stagnation pressure $(\rho/2) \cdot v^2$. The Bernoulli equation tells us that the sum of these three pressures is constant (under the conditions that have been mentioned).

Deficiencies:

Qualitatively and put into words, the equation states the following:

1. At places where the fluid is rapid, pressure is lower than where it is slow.
2. Pressure increases when going down within the fluid.

These statements contain only one pressure, which is the quantity p in the Bernoulli equation. Both terms $\rho \cdot g \cdot h$ and $(\rho/2) \cdot v^2$ have the dimension of a pressure, but they are not what we normally understand by a pressure. The terms of a sum do not necessarily represent physical quantities of the same kind. The term $\rho \cdot g \cdot h$ cannot be the gravitational pressure, since the gravitational pressure increases when going downwards whereas $\rho \cdot g \cdot h$ decreases.

Origin:

Probably the objectionable interpretation is due to the desire to consider pressure as a quantity for which a kind of conservation law is valid. Indeed, the formulation “The total pressure is constant” reminds a certain way of expressing the conservation or energy, electric charge or angular momentum: “In an isolated system the total amount of energy (electric charge, angular momentum) remains constant.“ Such statements are elegant, since it is easy to formulate them and they are universally valid. Thanks to the Bernoulli equation also pressure could enter the illustrious circle of the conserved quantities. Moreover, such a conclusion seems natural, since Bernoulli’s equation can be derived from the energy conservation law.

We believe that arguing in this way is exaggerating. Pressure cannot be a “conserved quantity”, since a necessary condition for being conserved is that the quantity is extensive – which is not true for the pressure.

One might object that, giving the name “dynamical pressure” to the term $(\rho/2) \cdot v^2$ can be neither false nor true, since it is only a question of giving a name. However, the choice of a name can be more or less appropriate and we believe, that the name pressure for the terms $\rho \cdot g \cdot h$ and $(\rho/2) \cdot v^2$ is not appropriate. Pressure is a quantity for which we have a sound intuition. Calling $(\rho/2) \cdot v^2$ a pressure would cause our students to believe that pressure is a difficult concept. The uplifting attribute “dynamic” further supports this idea.

Moreover, the expression $(\rho/2) \cdot v^2$ is known as the density of the kinetic energy. So one could argue to call all the terms of the Bernoulli equation energy densities: We might call p the static energy density, $(\rho/2) \cdot v^2$ the dynamic energy density and $\rho \cdot g \cdot h$ the gravitational energy density. It is obvious that this would not be a good idea.

Disposal:

Read the Bernoulli equation as follows: The pressure decreases when (1) the velocity increases, (2) height increases. Both statements are plausible.

Friedrich Herrmann

5.16 Force and energy

Subject:

Formulations containing the words “driving force” as the following

• “The driving force from the engine pushes the car along.”

• “Since the driving force of a car comes from the engine, it is…”

• “In a hybrid vehicle … a driving force transfer system is constituted…”.

• “The driving force, that is transmitted from the engine to the wheels…”.

• “A car that is running slowly can be accelerated by the force of the engine.”

Deficiencies:

If something is transmitted from A to B, then, according to common linguistic usage, it is first situated at A and then at B. Regarding the above citations, some of which are taken from schoolbooks, this does not apply if using the term “force” in the sense of physics, even when being generous. The statements are correct, however, when replacing the word “force” with what in physics is called energy.

Origin:

Nowadays, in physics the word “force” is used for the quantity **F**. But there is also a long tradition according to which the word had other meanings. So, in former times the word was used to describe what today we call energy – our kinetic energy was called *living force* or *vis viva*, as well as what we now call momentum. The famous historical dispute between Leibniz and the Cartesians about the “true measure of force”, which was about the question wether the expression $m \cdot v$ or $m \cdot v^2$ is the “true” measure for the description of a movement, attests to it. Apparently, the tradition is so strongly ingrained in the scientific terminology that even nowadays the word “force” is often taken for what should be called “energy” without being perceived by the author and the reader. Therefore we should not incriminate our pupils or students when they do not keep the concepts apart.

Disposal:

Careful wording when in physics it is about force.

Friedrich Herrmann

5.17 Pulleys

Subject:

"Like the lever, pulleys can also multiply force and change its direction."

"This pulley does not multiply the input force. It does change the direction of the force from up to down, and for many people, that is an advantage."

"A pulley changes the direction of the force, making it easier to lift things to high-rise areas."

Deficiencies:

A person A is pulling on a rope, which runs over a pulley B and on which a load C is hanging, Fig. 1.

1. The citations refer to forces in parts R_1 and R_2 of the same rope. If a force is mentioned without specifying the body that exerts it and that on which it is exerted, the orientation of the vector arrow is not yet defined.

The force $\vec{F}_{AB}$ that the person exerts an the pulley is oriented downwards, the force $\vec{F}_{BA}$ which the pulley exerts on the person points upwards. The state of the left part of the rope, i.e. R_1 is unambiguously described by the one or the other.

Our citations claim that the direction of a force is changed. This statement will be understood as follows: When going from part R_1 to R_2 of the rope, the force changes its direction. But we see that it is left to our discretion if it does so or not. $\vec{F}_{AB}$ has indeed the opposite direction of $\vec{F}_{BC}$ and thus a force seems to change direction. But the directions of $\vec{F}_{AB}$ and $\vec{F}_{CB}$ are the same and thus the direction of forces seems not to be changed.

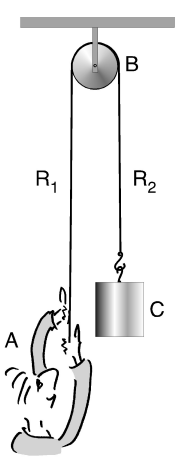


Fig. 1

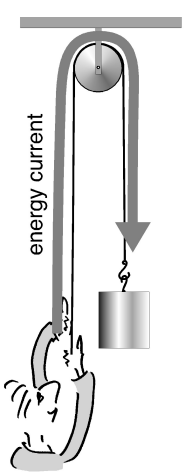


Fig. 2

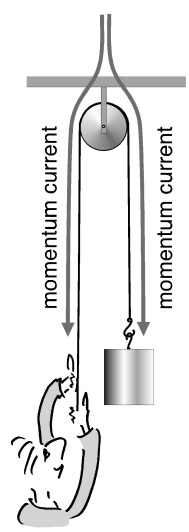


Fig. 3

2. "Change a direction" means that something that first has a certain direction later has another one. Thus, the sentences suggests that the force is going from rope 1 (left) to rope 2 (right). But if that would be so, what about the third force: that which the suspension exerts on the pulley? Where does this force go? Actually the pulley changes the direction of something, and even of two things: first it changes the direction of the rope, and second that of the energy flow, Fig. 2. When the rope R_1 is pulled downwards by the person, there is an energy flow upwards in R_1, it then goes around the pulley and down in part R_2 of the rope towards C.

3. It is indeed possible to handle the force in the way that is suggested by the citations. A force can be identified or interpreted as a flow of momentum. The momentum flow of our pulley arrangement can easily be given. However, momentum does not flow as one might suspect when reading our citations. It does not follow the rope around the pulley. If we take the upwards direction as the positive momentum direction, then momentum is flowing from the suspension into the pulley. There it branches into two currents of equal magnitude, one of them flowing through R_1 and the other through R_2, Fig. 3.

Origin:

The arrangement suggests a description with something that is flowing around the pulley in addition to the rope. But apparently the behavior of the energy flow is projected on that of the force, see also [1].

Disposal:

The idea of a force that changes direction is not much good. Things become clear when describing the pulley as well as the tackle with the flows of energy and momentum whereby one carefully keeps one apart from the other, just as in electricity one thoroughly has to distinguish between the energy flow and the flow of electric charge.

[1] *F. Herrmann:* Force and energy, article 5.16

Friedrich Herrmann

5.18 How an airplane flies

Subject:

How an airplane flies is discussed not only in the specialized literature about aerodynamics but also in physics text books for the University and the school and in the popular science literature. Various different explanations can be found. One source gives one explanation, another source another one, and some books give several descriptions. The most frequent explications are the following:

• The flow velocity is higher at the upper side of the wing than at the lower side. Therefore, according to Bernoulli's equation, the pressure is higher at the lower side than at the upper side.

• The molecules of the air are reflected at the surface of the wing. The momentum transfer is higher at the lower side than at the upper side.

• A circulation flow is forming around the wing. This causes a resulting force in the upward direction.

These explanations do not correspond to different mechanisms that contribute to the lift. They represent different descriptions of the same phenomenon.

Deficiencies:

1. Often the explanations are not understandable. Some texts inundate the reader with details and technical terms: boundary layer separation, hull resistance, Reynolds number, angle of attack, lift coefficient, circulation, vorticity, viscosity, Stokes' law, Bernoulli equation, turbulence…

2. Some texts suggest that the above mentioned processes correspond to various contributions to the lift. In one book it is said that the molecules transmit momentum to the lower side of the wing, and that in addition there is an underpressure at the upper side.

3. The most important problem that we want to discuss here is: When a pupil asks, why an airplane flies, what kind of answer will satisfy him or her?

We believe that none of the above-mentioned explanations represents such an answer. In order to see why, let us consider another question that has something in common with the flying airplane. Instead of: "Why does the airplane not fall to the ground?" we will ask "Why does the vase on the table in front of us not fall down?"

A hypothetical physicist, who argues with Bernoulli's equation in the case of the airplane, might consequently give an answer like the following: "The table is elastic. It behaves like a tended spring and thus exerts a force on the vase." Obviously, this answer is not incorrect, but probably the questioner is not really interested in who exerts a force for which reason on the lower side of the vase. The same is true for the airplane. If we know how the air manages it to push the wing upwards, we do not feel to have understood the reason of why the airplane does not fall down.

That hypothetical physicist who argues with the air molecules in order to explain why the airplane flies, would, in the case the vase answer: "The vase does not fall to the ground, because there is a repulsive interaction between the molecules of the vase and those of the table." With this answer the questioner would not be very happy either. And also the corresponding answer in the case of the airplane does not bring the desired insight. Do we really have to haul out atomic physics to understand why the airplane flies?

Perhaps our hypothetical physicist did not really listen to the question. Apparently he translated the pupil's question into another one which allows him to display all his knowledge about aerodynamics and molecular kinetics.

Origin:

What we find in school books and in other popular science books, i.e. books that are written for non-specialists, is unprocessed engineer's knowledge. It is important to know the streamline field when the problem is to optimize an airfoil. It is important to decompose the velocity field in components with vanishing divergence and vanishing circulation respectively in order to apply potential theory. But these subjects cannot pretend to be themes of a general education.

Disposal:

We confine ourselves to the following explanation: Just as a bird, an insect, a helicopter, a frisbee, a boomerang and also a parachute, an airplane must set the air into a downward motion since it must get rid of the momentum that it gets steadily by means of the gravitational force. (Actually, force is identical with momentum current.) The air that is going downwards takes this momentum away and eventually brings it back to earth.

Friedrich Herrmann

5.19 Angular momentum conservation

Subject:

The law of conservation of angular momentum is often introduced as follows: We consider a mass point. We write the cross product of the vector on both sides of Newton's second law

$\boldsymbol{F} = \mathrm{d}\boldsymbol{p}/\mathrm{d}t$

and the position vector $\boldsymbol{r}$ (relative to an arbitrarily chosen origin). We get a relation between the torque and the time rate of change of the angular momentum:

$\boldsymbol{M} = \mathrm{d}\boldsymbol{L}/\mathrm{d}t$.

We write the corresponding expression for two or more mass points and take into account that for the internal interaction forces there is

$\boldsymbol{F}_{ik} = -\boldsymbol{F}_{ki}$,

and that these forces are parallel to $\boldsymbol{r}_i - \boldsymbol{r}_k$. We then find that the time derivative of the angular momentum of the system of mass points is equal to the sum of the torques of the external forces. From this follows the law of angular momentum conservation: "The angular momentum of a system remains constant, if no external torque acts on the system."

Deficiencies:

Our foregoing derivation of the angular momentum conservation is somewhat short, since we believe that it is known to the reader. In a text book it easily needs an entire page with about 10 lines of equations. It is not hard to follow such a derivation step by step, and at the end, the student will probably be convinced that the law of angular momentum conservation must be valid. However if we ask the student, what actually has been proven on this page, he or she might run into trouble. The derivation starts with Newton's second law, which is equivalent to the law of momentum conservation, and the result of the calculation is the law of angular momentum conservation. It is unavoidable that the student believes, angular momentum conservation has been mathematically derived from momentum conservation. It is needless to say that this is not true. There will hardly be a student who understands the trick that was employed. They will even not suspect that there was a trick. Actually, in the above derivation angular momentum conservation is not derived from momentum conservation, but angular momentum conservation is fed into the calculus when saying that the forces $\boldsymbol{F}_{ik}$ and $\boldsymbol{F}_{ki}$ are parallel to $\boldsymbol{r}_i - \boldsymbol{r}_k$.

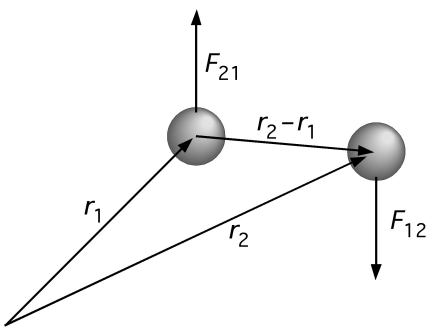


Fig. 1. The forces on both bodies are not parallel to the straight line between the bodies. Thus, the law of angular momentum conservation is not obeyed.

Fig. 1 shows something that does not exist in reality. Two bodies exert forces on one another, that are equal and of opposite direction ($\boldsymbol{F}_{12} = -\boldsymbol{F}_{21}$), but which are not parallel to $\boldsymbol{r}_i - \boldsymbol{r}_k$. They obey Newton's third law and thus momentum conservation, but since they are equivalent to a torque, the angular momentum of the system should increase, and it would do so without an external torque. But there are no such forces. They are forbidden by the law of angular momentum conservation. Thus the claim that $\boldsymbol{F}_{ik}$ and $\boldsymbol{F}_{ki}$ are parallel to $\boldsymbol{r}_i - \boldsymbol{r}_k$ is equivalent to the claim that angular momentum is a conserved quantity.

In summary, a somewhat lengthy calculation is carried out, into which angular momentum conservation is injected, and at the end, one is happy that the law of angular momentum conservation comes out. But why then the calculation?

Origin:

Newton's laws contain not more and not less than momentum conservation. Due to their great success the idea has spread that they are more than just a simple conservation law. They seem to be the be-all and the end-all of physics, the basis from which everything else can be derived. Sometimes, even energy conservation is derived from Newton's laws – again with a trick.

Disposal:

Introduce angular momentum as a quantity of its own right, for which a conservation law is valid. This does not exclude to show how the angular momentum of a system of mass points is related to the momenta of its constituents.

5.20 Inertial frames of reference

Subject:

The concept „inertial frame of reference“ plays an important role in the teaching of physics at school and at university. It is needed to formulate the law of inertia:

> „A body either remains at rest or continues to move at a constant velocity, unless acted upon by an external force.“

This law only holds as long as the corresponding process is described in an appropriate reference frame: an inertial frame of reference.

And what is an inertial frame of reference?

> „In physics, an inertial frame of reference (from latin *iners* „idle, languid“) is a coordinate system, in which a body moves at constant velocity unless a force is acting on it.“

An inertial frame of reference can be „realized“ approximately:

> „A nearly perfect approximation to an inertial references frame is realized by a spaceship that moves through interstellar space, far from all masses, as long as it is not rotating.“

Deficiencies:

Let us first have a look at the law of inertia:

> „The velocity of a body remains constant unless an external force is applied to it.“

Taken alone, this statement cannot be true, since any body moves with constant velocity, if the frame of reference is chosen appropriately. Thus, the law of inertia cannot be valid generally. It is valid, so we learn (sometimes afterwards), only in certain frames of reference, the so-called inertial frames. We thus can better formulate the law of inertia:

> „In an inertial frame of reference the velocity of a body remains constant unless an external force is applied to it.“

But how do we know whether a frame of reference is inertial or not? One considers a body of which it is known that no force is acting on it as a reference body. This reference body defines the inertial frame of reference. We can then decide for each other body whether it moves with constant velocity or not. So far, so good.

However, one question remains: How do we recognize that no force is acting on our reference body? We cannot say: Because it moves with constant velocity – because it is just this reference body that tells us what is meant by constant velocity. We therefore have to decide by another method whether no force acts on the body. At a first glance, this seems to be simple. We know the forces of nature; we know the sources and we know the laws that govern the dependance of the force on the distance. We thus can, at least in a thought experiment, make sure, that no force is acting: we can ensure that no electric, or magnetic, or contact forces are acting.

We also have to make sure that there are no gravitational forces. But here comes the problem. We cannot exclude gravitational forces. As long as one believed that gravitostatic forces (i.e. forces that can be calculated by means of Newton's law of gravitation) can be distinguished from inertial forces, there was no problem – see above: It was sufficient to go to the interstellar space. However, if we admit, that these two types of forces cannot be distinguished in principle, the argument fades away. We no longer can decide whether a force is acting on a body or not. A force is acting in one frame of reference and not in another.

Imagine we are in a spaceship in interstellar space and the rocket engine works at constant thrust. Is the law of inertia valid in the spaceship or not? Of course it is valid. If we release a body or if we throw it away, it does not move on a straight line, which is normal since a force is acting on it. In the spaceship, taken as our reference system, the gravitational field strength is not zero, we thus have a gravitational force acting on the body, and the body, in accordance with the Second law is accelerated. Consequently, Newton's laws are also valid in this „non-inertial reference“ frame.

Why then are we using the concept of an inertial frame of reference?

Origin:

The reason why the reference frame, for which Newton's laws supposedly are valid, is often not mentioned, may be that because Newton himself does not mention it in his formulation of the First and Second law. Newton does not need to do so, since he explains the conditions for the validity in the „Definitions“ that precede the chapter with the „laws“: There he explains what he means by true forces and by fictitious forces. A true force can be recognized by the fact that there is another body that exerts the force. This is not true for fictitious or inertial forces. His laws are valid for true forces only.

Already twenty years after the publication of the *Principia* this idea was questioned. The Irish Philosopher George Berkeley (1685 - 1753) proposed, that also the inertial forces have their origin in other bodies, namely in the fixed stars. This idea was taken up later by the physicist and philosopher Ernst Mach. According to this point of view the criterium for a fictitious force was not longer fulfilled. Inertial forces had become true forces.

Within the physics of gravitation for a long time a strange situation persisted: It was known that inertial and gravitational mass are equal, but there was no explanation for it. In 1916 Einstein commented this fact as follows:

> „The gravitational and the inertial mass of a body are equal. The hitherto mechanics has registered this important fact, but not interpreted.” [1]

General relativity tells us, that a phenomenon or process, that is explained in one reference frame by means of inertia, is described in another frame by gravity. In this way the distinction between gravitational and inertial mass disappears, just as the distinction between an inertial and a non-inertial reference frame.

In 1922 Einstein writes:

> „The genuine achievement of the (general) theory of relativity is that it frees physics from the necessity of the introduction of the „inertial reference frame“ (or inertial reference frames). The unsatisfactory of the concept is: It picks out without justification certain systems among all thinkable coordinate systems. It is then supposed, that the laws of physics are valid *only* for these inertial systems (e.g. the law of inertia and the law of the constancy of the velocity of light). In this way one attributes a role to space that distinguishes it from the remaining elements of the physical description: It acts in a determining way on all physical processes, without these influencing the space; although such a theory is logically possible, it is rather unsatisfactory. Newton had clearly perceived this shortage, but he also had understood, that for the physics of this time there was no other way. Among the later scientists, it was particularly Ernst Mach, who brought this point to light.“ [2]

Disposal:

One might believe that one can do without the concept of an inertial frame only if gravitation is treated in the context of general relativity, i.e. that general relativity must be known with its tensor calculus, with Einstein's field equations, with the Ricci tensor etc. But this demand would be exaggerated. The identity of gravitational and inertial mass is a consequence of the theory of general relativity, but it can also be understood without it. All we have to do is to admit that the identity of the two masses is not accidental.

Here a short sketch of how the subject can be treated in the classroom. Consider the situation of Fig. 1. Willy is in the famous falling elevator; Lilly is outside. For Willy (his reference frame is the elevator) the globe does not move; it is floating. He thinks this is normal, since in his reference frame the strength of the gravitational field *g* is zero, no force acts on the globe. Lilly's interpretation is different: the field strength is not zero, a force acts on the globe.

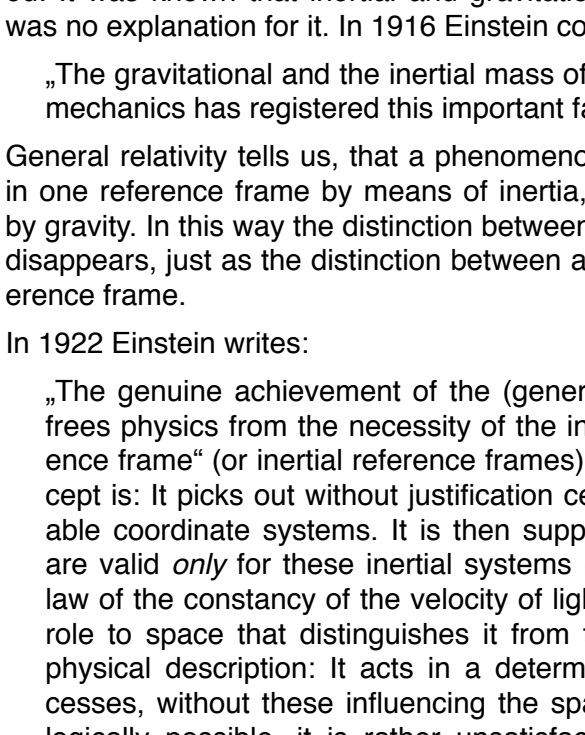


Fig. 1. Willy: „The globe is floating, the gravitational field strength must be zero.“ Lilly: „The globe it falling, its velocity increases. The field strength is not zero.“

We conclude, that the value of the gravitational field strength depends on the reference frame. We are not surprised because a similar behavior is known for many other physical quantities.

In fig. 2, the question is why the spring is stretched. In Willy's opinion (in his reference frame), the gravitational field strength is zero everywhere. The spring is stretched, because the body that is attached to it, is accelerated. Due to its inertia or its inertial mass it resists to an acceleration.

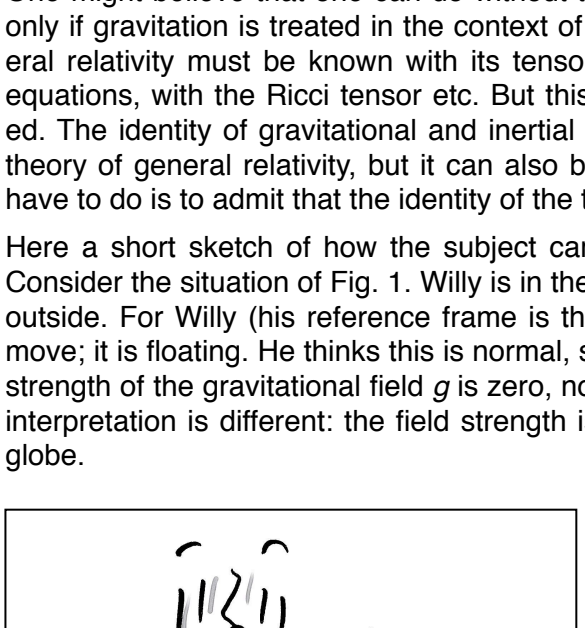


Fig. 2. Willy: „The field strength is zero. The spring is stretched because the body, which is accelerated has inertia.“ Lilly: „The spring is stretched because the body is pulling at it due to its weight.“

Lilly on the contrary is convinced that the stretching of the spring has nothing to do with its inertia, but is due to the gravitational force exerted on the body, and thus to the gravitational mass of the body.

We conclude: According to the reference frame mass manifests itself either as gravity or as inertia. We see that the distinction between gravitational and inertial mass, just as the distinction between true forces and fictitious forces or between inertial and non-inertial reference frames is an artefact of prerelativistic physics.

[1] *A. Einstein:* Über die spezielle und die allgemeine Relativitätstheorie, Akademie-Verlag Berlin (1973), S. 54.

[2] *A. Einstein:* Grundzüge der Relativitätstheorie, Akademie-Verlag Berlin (1970),S. 138.

Friedrich Herrmann

5.21 Tug-of-war

Subject:

Two persons A and B compete in tug-of-war. By means of a rope A exerts a force on B, and B exerts a force on A, Fig. 1.

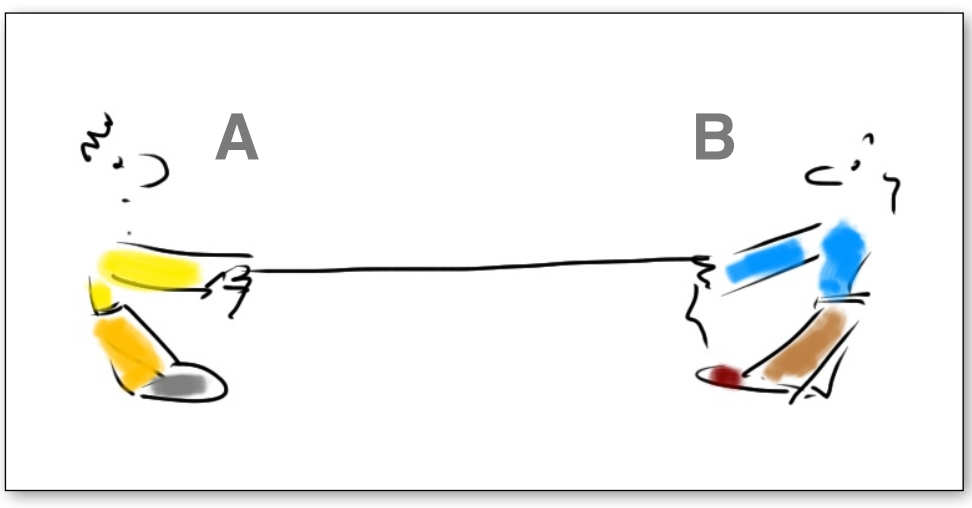


Fig. 1. A exerts a force an B, and B exerts a force on A.

Deficiencies:

In a seminar 17 students who study to become physics teachers (3rd or 4th year) are asked to sketch the forces in a figure with two persons playing tug-of-war, and discuss the relations between these forces. They work in small groups and are allowed to discuss with one another. No help is given by the professor. They are asked to present their results.

They begin with the simplest case: Both persons A and B are at rest and remain at rest, thus: velocity zero and acceleration zero. They had been informed, that each of the persons “pulls with a force of 200 N”.

It turns out that there are three different opinions among the students about which force “acts within the rope”.

Opinion 1: The force in the rope is zero, since (+200 N) + (–200 N) = 0;

Opinion 2: The force is 400 N, since 2 · 200 N = 400 N;

Opinion 3: 200 N, as the persons pull with 200 N.

Since in the discussion they do not come to an agreement they decide to vote. Surprisingly, now all of them vote for 400 N.

Because there is still an uneasiness about the claim, they discuss how one might decide about the correct answer with an experiment. They agree about the proposal to insert three spring scales in the rope, Fig. 2: one next to A to measure the force exerted by A, one on B’s side to measure the force of B, and one in the middle to measure the force “in the rope”.

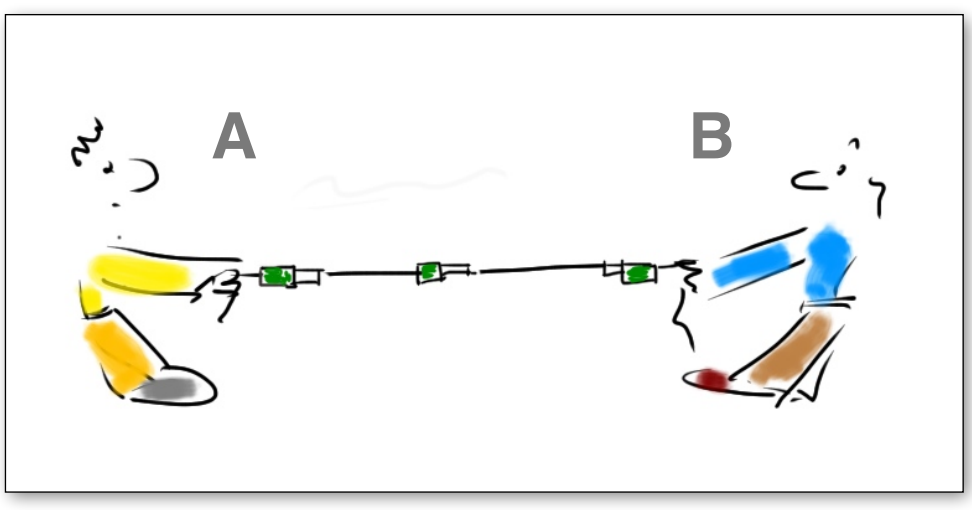


Fig. 2. “The spring scale at the left measures A’s force, that on the right measures the force of B. That in the middle measures the force in the rope.”

Since the physics lab is next door, the professor proposes to really make the experiment – what indeed is done. The students are surprised about the result.

This happened recently at the physics faculty of the Karlsruhe Institute of Technology. At that time the 17 students had “learned” Newtonian mechanics already three times: at the lower secondary school, at the upper secondary school and at the University. In addition they had learned Hamiltonian and Langrangian mechanics. Moreover, they did not give the impression to be particularly untalented or unintelligent.

We believe that it is not exaggerated to classify this result as a fiasco. We could report about similar experiences with other problems of elementary mechanics. We conclude: Students do not understand Newtonian mechanics.

Origin:

For sure, the students cannot be blamed for this embarrassing result and also not necessarily the teachers and professors. At least we cannot accuse the teachers not to understand the physics they are teaching. The result seems to be independent of who was the teacher of our students. Thus, we better not ask for the fault, but for the causes, and the causes are easy to find: It is the Newtonian way of describing a momentum transfer.

Consider the situation depicted in Fig. 3: Two bodies A and B are connected by a tended spring. The momentum of A increases, that of B decreases.

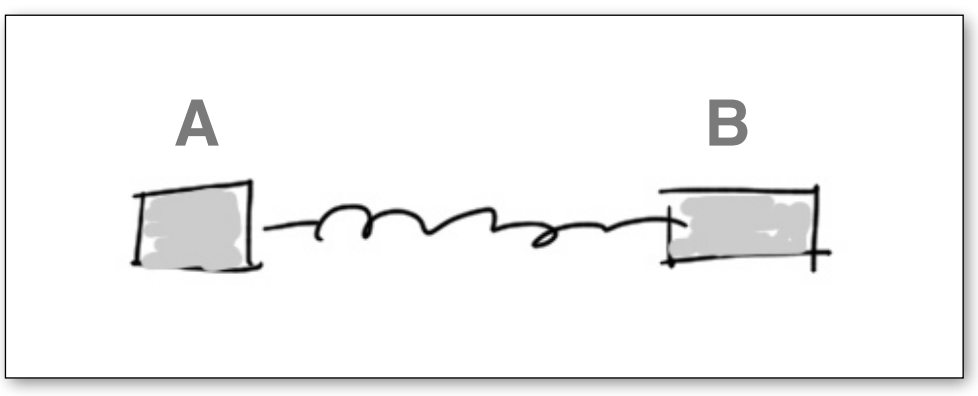


Fig. 3. The spring is under tensile stress. There is a momentum current from right to left.

If we take the local conservation of momentum seriously, we have to admit that momentum proceeds through the spring from B to A, or in other words: In the spring a momentum current is flowing from B to A. If the intensity of the current as it leaves B is known, one also knows it at the position where it enters A, and also at all the positions between A and B. More exactly: at every section through the connection between A and B.

We have just described the situation with a model that in other fields of physics has proved to be useful: the substance model. Momentum is imagined as a kind of substance or fluid, in the same way as we do with mass or electric charge. The change of momentum of a system can take place only by an in- or outflow of momentum.

Newton could not use or introduce this simple model, since to do so, one needs the field concept. But fields were yet unknown at Newton’s time. The most important bodies for which to apply his laws were the celestial bodies. How does momentum (the *quantitas motus*) get from the Earth to the Moon, or from the Moon to the Earth?

Newton did not know enough about a system localized between Earth and Moon, that today we call gravitational field. Therefore his “Hypotheses non fingo”. And therefore his somewhat unwieldy force model. Instead of saying “momentum goes for A to B” we have to say: “A exerts a force on B, and B exerts a force on A and thereby the momentum of both bodies changes.” It was clear to Newton that this could not be the final word on the matter [1]. Newton’s force model allows for a coherent description of a mechanical interaction, but its handling is difficult, as we have seen.

Disposal:

Today we are in a more comfortable situation than Newton was. We can confidently suppose that Newton, had he known the field concept, would have described the processes that we are interested in with momentum currents.

Friedrich Herrmann

[1] Letter of Newton to Richard Bentley; The Newton Project
http://www.newtonproject.sussex.ac.uk/view/texts/normalized/THEM00258
„The last clause of your second Position I like very well. Tis unconceivable that inanimate brute matter should (without the mediation of something else which is not material) operate upon & affect other matter without mutual contact; as it must if gravitation in the sense of Epicurus be essential & inherent in it. And this is one reason why I desired you would not ascribe innate gravity to me. That gravity should be innate inherent & essential to matter so that one body may act upon another at a distance through a vacuum without the mediation of any thing else by & through which their action or force may be conveyed from one to another is to me so great an absurdity that I beleive no man who has in philosophical matters any competent faculty of thinking can ever fall into it. Gravity must be caused by an agent acting constantly according to certain laws, but whether this agent be material or immaterial is a question I have left to the consideration of my readers.“

5.22 The direction of momentum currents

Subject:

A force can be interpreted as the intensity of a momentum current, or as a momentum current for short. This understanding is due to Max Planck [1], and it is mentioned or employed in various text books of Theoretical Physics, in particular in the context of the mechanics of continuous media. Since momentum is a vector quantity and since its cartesian components can admit positive and negative values, there is an arbitrariness in the choice of the sign. The direction of the flow of the three components of momentum also depends on this choice.

Deficiencies:

Since momentum currents are treated only in text books at an advanced level, most physics students do not learn about this interpretation. As a consequence, even experienced physicists feel unsure in handling this concept. Some feel uneasy about the fact that the direction of the current depends on the arbitrary definition of the sign of the quantity that is flowing [2].

In physics the direction of the "flow of a physical quantity" is always the direction of the current density vector, whatever the flowing quantity may be.

Since the problem of the direction of a current is the same in mechanics as in electricity, let us first remind, how the direction of an electric current is defined.

For the electric charge a continuity equation holds:

$$\frac{\partial \rho_Q}{\partial t} + \mathrm{div}\, \vec{j}_Q = 0 \qquad (1)$$

Here, ρ_Q is the electric charge density and $\vec{j}_Q$ the electric current density.

If the charge density decreases at a given point, i.e. if $d\rho_Q/dt < 0$, the divergence of $\vec{j}_Q$ is positive, charge flows away from the point. This definition of the direction of a current follows from the equation of continuity for the electric charge. It is not based on a convention, as is sometimes said.

However, there is a possibility to obtain the opposite direction for the electric current, namely by redefining the sign of the electric charge. According to a general agreement, electrons carry negative, protons positive charge. If we define the charge of the electrons as positive and that of the protons as negative, equation (1) would tell us, that the current is flowing in the opposite direction. Thus, we can say that the direction of the electric current is based on a convention, but not in the way that some text books assert: We cannot maintain the convention about the sign of the electric charge and flip only the direction of the current.

Back to momentum: Each of the three cartesian components of momentum obeys a conservation law and, as a consequence, for each of them a continuity equation can be formulated, similar to equation (1). Just as in the case of the electric charge, we have (for each component) the freedom to define what we will understand by positive and negative momentum. As soon as we have decided about this sign, the direction of the current is also defined. And when we change our mind and define the positive momentum the other way round, the current direction will flip.

There is no problem in finding an agreement about the definition of the sign of electric charge. When the decision is taken it is mandatory. Not so with momentum. For every new experiment or concrete situation the decision has to be taken anew. Actually, there is a convention, but it does not help very much: If a body moves to the right, the *x* components of its velocity and momentum are positive. This convention comes from mathematics and has been adopted by the physicists: the positive direction of the horizontal axis of any coordinate system is to the right. However, it suffices to observe a movement from behind to get a disagreement: For the observer who looks from behind a velocity component that is positive for us, will be negative for him. As a teacher we are often in this situation. Therefore, an experiment on the teacher's table is best described from the view point of the students.

Origin:

The change of the current direction upon a change of the definition of the positive momentum direction may appear unaesthetic. Who is not familiar with momentum currents, may complain about a seeming symmetry break in a process or phenomenon that obviously is symmetric. Fig. 1a shows a simple momentum current circuit. The spring is under tensile stress. Momentum is flowing counterclockwise. (In addition to the sketched current, other currents are flowing within the rigid yoke.) If we now define the direction of positive *x* momentum to the left, the momentum current changes its direction, Fig. 1b.

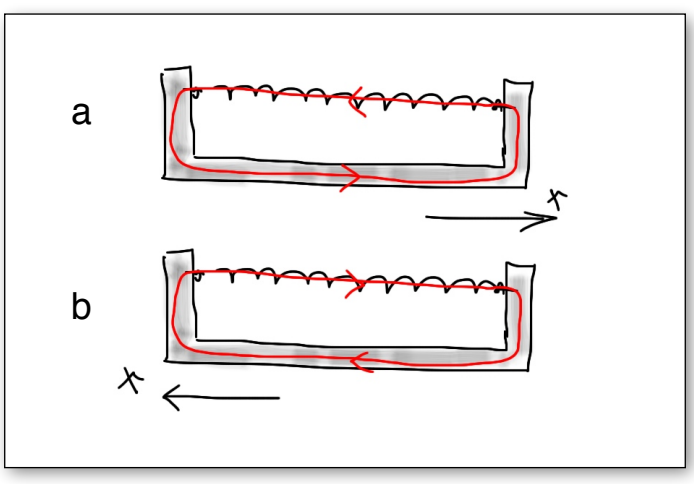


Fig. 1. (a) Positive momentum (movement) to the right: momentum flows counterclockwise; (b) positive momentum to the left: momentum flows clockwise.

However, it is surprising that experienced physicists take offense at this observation. This kind of symmetry breaking is always the price to be paid when a problem is described mathematically. As soon as we choose a coordinate system we break a symmetry between left and right, upwards and downwards etc. When calculating electronic orbitals, all of a sudden a preferential direction, usually called *z* direction, appears in a situation of spherical symmetry. Each beginner has a problem with this fact, but eventually he understands that the distinction is caused only by the mathematical description.

We also know from other situations that a current changes its direction solely upon changing the reference frame. Consider the energy flow in a bicycle chain. In the reference frame of the bicycle the energy goes through the strained part of the chain from the driving sprocket to the driven sprocket. In a reference frame in which the tended part of the chain is at rest (which moves in the travel direction of the bicycle, faster than the bicycle itself), there is no energy flow at all. And in a reference system that moves even faster (relative to the Earth), the energy reverses its direction: it flows from the driven sprocket to the driving sprocket. (Don't be afraid that the net energy flow to the back wheel has changed by a mere change of the reference system; energy is also flowing through the bicycle's frame, as soon as the frame is moving.

Would the "experts" of the German Physical Society also in this case claim, that "such a current does not exist in nature", or that "it is not a property of the system"?

Disposal:

When introducing electric currents at school do not focus on the movement of the charge carriers, but employ right from the beginning the "substance model": We imagine electric charge as a stuff that can flow in an electric conductor. The flow direction follows from the balance equation. At school it is even not necessary to formulate the continuity equation. It is a matter of course to say, that electric charge flows away from a body when the charge of the body decreases.

In the same way we deal with momentum. Here, it is important to make clear from the beginning which direction we choose as the positive momentum direction.

Friedrich Herrmann

5.23 Momentum currents in momentum conductors at rest

Subject:

It is well-known that momentum is transferred by a moving body or by a flowing liquid or gas. These processes are sometimes called convective momentum currents. It is less well-known or even disputed that momentum also flows in a medium at rest, as long as it is under mechanical stress [1]. The corresponding current is sometimes called a conductive momentum current.

Deficiencies:

Fig. 1 shows a somewhat unusual oscillator. Two bodies A and B that can move horizontally, are coupled by two springs and a bar [2]. We suppose that the bar is absolutely rigid and that the springs are massless* (as one often does in mechanics).

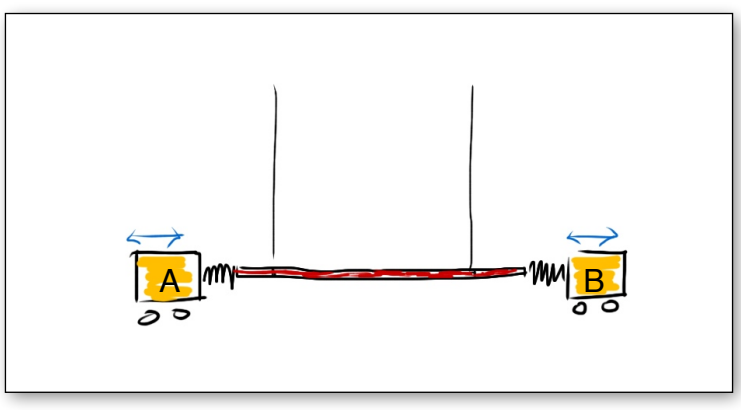


Fig. 1. A and B move in opposite directions. Thereby the bar remains at rest.

Now the two bodies are displaced outwards and then released, so that they begin to oscillate. The momenta of A and B change periodically in such a way that the decrease of the momentum of one of them goes together with an increase of equal magnitude of the momentum of the other. We can express this fact in other words: Momentum is going back and forth between the two bodies.

If one does not apply the momentum flow interpretation for the quantity ***F***, one cannot use this simple description. Instead one has to say: A exerts a force on the left spring, and this spring exerts a force on A. Moreover, the left spring exerts a force on the bar, and the bar on the left spring. In addition the bar exerts a force on the right spring and the right spring exerts a force on the bar. Finally the right spring also exerts a force on body B, whereas B exerts a force on the right spring. Here, we did not even say anything about what is the relation between all these forces.

Origin:

We suppose, that the low acceptance of non-convective momentum currents is due to a somewhat naive idea about the concept of a current in physics. According to this point of view there is a current only if there is a movement of a substance or a collective movement of particles. With such a concept of a current it would be logic to say there is no current when there is no moving substance or when there are no moving particles. However, this is not the concept of a current or a flow neither in the colloquial speech nor in physics.

In the everyday language we also speak about currents, when not referring to the movement of some material entity, but to something that in physics would be called an extensive quantity: We are used to speak about a flow of money or a flow of data.

In physics, things are even simpler. We say there is a current of the quantity *X* when an equation of the following form can be formulated:

$$\frac{\partial \rho_X}{\partial t} + \mathrm{div}\, \vec{j}_X = \sigma_X$$

This equation is called equation of continuity and can be interpreted as a balance equation or an accounting equation. The names that are given to the physical quantities in the equation are due to this interpretation: ρ_X is called the density of *X*, $\vec{j}_X$ is the current density and σ_X the density of the production rate (which is zero if *X* is a conserved quantity).

This equation does not require that ρ_X be non-zero at all points where , $\vec{j}_X$ is non-zero [3].

Actually, this case, $\vec{j}_X \neq 0$ whereas $\rho_X = 0$) can be realized, whenever the flowing quantity *X* can admit positive as well as negative values. Then it is allowed to imagine the actual current to be the result of two contributions where the density of these contributions add up to zero, whereas the current densities do not.

An example is an electric current in a normal electric conductor. The electric charge of the positive and negative charge carriers add up to zero, whereas the corresponding current densities do not.

This possibility does not exist for extensive quantities that admit only positive values, such as energy, mass, entropy or amount of substance.

Once more: the validity of an equation of continuity is the only justification for a physicist to use this model. No moving particles are required.

Who believes to better understand a current if it is coupled to the movement of particles may consider the momentum transfer by means of a non-moving gas, Fig. 2. Both pistons are accelerated outwards by the gas; piston A to the left, piston B to the right hand side. Obviously momentum gets from left to right (we have defined: momentum is positive when the movement is to the right). In this case the microscopic mechanism of momentum transfer is so obvious that there will hardly be any doubt about interpreting the process as a momentum flow.

Those molecules whose velocity has a positive *x* component, carry positive momentum to the right. The molecules with a negative *x* velocity have negative momentum and they carry it to the left – which also corresponds to a transfer of (positive) momentum to the right. We see that the contributions of the two classes of molecules to the total *momentum density* cancel, whereas their contributions to the *current density* add up constructively.

A problem that one may see is that the direction of flow changes upon changing the definition of the positive momentum direction. Our example shows that there is no mystery behind.

A similar reasoning can be applied to the momentum flow through a solid body or through the electromagnetic field. Then, things are somewhat more intricate**. But there is no new insight relative to the difficulties that are seen by the members of the board of the German Physical Society. Things are as simple as they appear when considering the gas, and that means: they can be understood by pupils of the lower secondary school.

Disposal:

Physical quantities are variables in the sense of mathematics. Therefore, they cannot flow as a matter of principal (just as mass cannot hang on a spring). If one nevertheless speaks of a current of electric charge, mass or momentum, it means that one is using a model. Who is aware of this fact, will not ask the question of whether there are momentum currents in nature. One may introduce them or one may not; one uses the model or one does not use it. A decision against the model of a momentum current would lead to the question of why one does use it in the case of energy, mass and electric charge.

Friedrich Herrmann

*This is justified for our purposes. Actually the assumption would have as a consequence that momentum is not propagating with the velocity of sound but with infinite velocity.

http://www.physikdidaktik.uni-karlsruhe.de/kpk/Fragen_Kritik/KPK-DPG%20controversy/Expert_opinion_english.pdf

*This condition is acceptable for our purposes. Actually, it would mean that the velocity of propagation of the momentum is not the velocity of sound, as it should be, but it would be infinite.

**We recommend as an exercise for the handling of momentum currents to make the corresponding reasonings for a thermally excited linear chain.

[1] Expert opinion on the Karlsruhe Physics Course; Commissioned by the German Physical Society; M. Bartelmann, F. Bühler, S. Großmann, W. Herzog, J. Hüfner, R. Lehn, R. Löhken, K. Meier, D. Meschede, P. Reineker, M. Tolan, J. Wambach und W. Weber;

http://www.physikdidaktik.uni-karlsruhe.de/kpk/Fragen_Kritik/KPK-DPG%20controversy/Expert_opinion_english.pdf

[2] I am grateful to Werner Maurer for the idea of this experiment, see also:

http://www.youtube.com/watch?v=aBLPEOM7xbM

[3] *Gustav Mie:* Entwurf einer allgemeinen Theorie der Energieübertragung, Sitzungsberichte der Mathematisch-Naturwissenschaftlichen Classe der Kaiserlichen Akademie der Wissenschaften, CVII. Band, Abtheilung II.a, 1898, S. 1113-1181

5.24 Direction of momentum current and coordinate system

Subject:

Two bodies A and B are connected by a strained spring, Fig. 1.

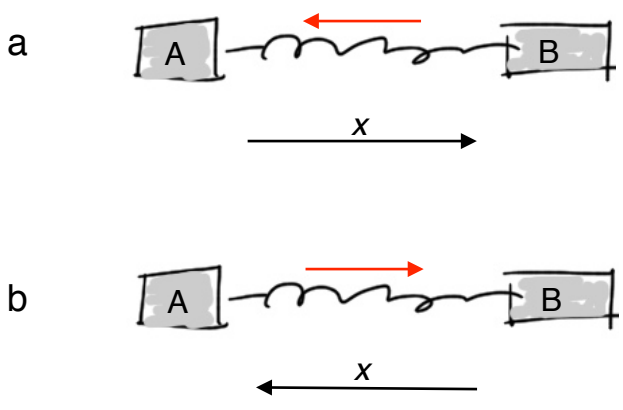


Fig. 1. The spring is strained; the red arrow points in the direction of the flow of *x* momentum. (a) The momentum of A increases, that of B decreases. (b) The momentum of B increases, that of A decreases.

In the upper image (a) of Fig. 1 the *x* axis points to the right. Therefore, the momentum of A increases whereas that of B decreases. We conclude that momentum is flowing from right to left, i. e. in the negative *x* direction, as shown by the red arrow.

If the *x* axis points to the left, lower image (b) the momentum of B increases and that of A decreases, i. e. momentum goes from left to right.

This cannot be true, because it would mean that we have „changed the direction of the momentum flow arbitrarily, i. e. independently of what is happening within the system, only by a new choice of the coordinate system“ [1].

Deficiencies:

The above reasonings contain an error.

But first a general remark.

A physical quantity describes a property of a system. The value of the quantity depends on several factors. In the first place it depends on the state of the system, since it is this state that we want to describe. It is trivial that it also depends on the measuring unit that had been chosen. Finally, it often depends on the choice of the reference frame. The fact that the values of physical quantities depend on the reference system is known to every physicist.

In the present case the apparent contradiction, or, more exactly the uneasiness, can be eliminated in two ways.

1. The flow direction of a current is, in mathematical terms, the direction of the current density vector. A vector is a representation of a quantity that is independent of the coordinate system. Velocity is a vector. If we represent it by an arrow, the direction of the arrow is independent of the choice of the coordinate system. The same is not true for the Cartesian coordinates. Consider two cars that are running both with a speed of 60 km/h in opposite directions, one to the left, the other to the right. To describe the situation physically, we attribute a vector to each car. This description has a „left-right symmetry“. If we employ Cartesian coordinates, we attribute to one of the cars a positive velocity +60 km/h and to the other a negative velocity –60 km/h. Now the symmetry is broken – a fact at which no physicist will take offense.

Shouldn’t it be the same with regard to the momentum flow? Isn’t the momentum current density a vector, so that its direction is independent of the choice of the coordinate system? No, the momentum current density is not a vector; it is a tensor. When we here refer to the direction of a momentum current, we refer to the current of one component of the momentum vector and this is not independent of the coordinate system. It depends on the coordinate system for the same reason as the component of the velocity did in our previous example.

2. Regarding the teaching at school, it is not necessary to know the concept of a tensor. We treat the three components of the momentum independently, as if we had to do with three scalars. For each of them taken alone a conservation law holds. Let us consider the *x* component of momentum. We not only have to choose the direction of the *x* axis, but also its orientation. If the *x* axis is horizontal, we can define positive momentum to correspond to a movement to the right or to the left. A body with a momentum of 5 units in the case of the first choice (positive momentum to the right) has –5 units in the second case (positive momentum to the left). Thus, our description of the system does not reflect the intrinsic symmetry of the system. But this is the sacrifice we have always to make when describing a system in cartesian (or cylindrical or spherical) coordinates. By choosing a coordinate system we destroy the symmetry of the system.

Actually, it is possible to formulate the rule for the flow direction of momentum in our spring in an invariant manner:

If the spring is under tensional strain (positive) momentum flows in the negative direction.

The statement remains correct when the definition of the positive momentum direction is changed. However, we do not recommend to use this formulation at school. Pupils do not have the same kind of difficulties as experienced physicists.

Origin:

The „experts“ of the German Physical Society seem to take offence at the fact that the mathematical description of a symmetric situation is unsymmetric. The reason may simply be that they never had to do with our particular situation. When you look around, you easily find many other phenomena where a similar problem arises, but which are not considered as difficulties – simply because one has got acquainted to the situation.

Here some examples:

When treating the hydrogen atom with quantum mechanics, the *z* direction seems to play a particular role. Each student has a problem with this fact and some of the students never understand that this *z* direction is only an artefact of the mathematical description.

Only few students get aware of what is involved in the famous two basic experiments of electromagnetic induction: In one of them (Magnet moving, coil at rest) the induced emf is caused because $\int \vec{E}_{ind} \cdot d\vec{s}$ is different from zero; in the other (Coil moving, magnet at rest), because there is a Lorentz force. In the first case the electric field is a curl field, in the second not. Actually both describe the same effect, but they are described in different reference frames. Here again, one might wonder that „independently of what is happening within the system“ two very distinct explanations are given for the same process.

With the same argument one might wonder why the energy flow in a bicycle chain sometimes flows backwards from the front sprocket wheel to the rear sprocket wheel and sometimes the other way round; in other words: the energy flow density vector points to the left or to the right, depending on the choice of the reference frame, i.e. „independently of what is happening within the system“. Would one conclude that the direction of the energy flow density vector does not describe „a property of the system“?

Or consider the magnetic field of a straight wire, through which an electric current is flowing. The field is caused by the (drift) movement of the mobile charge carriers, i.e. the electrons – that is what is usually said. However, in the reference frame in which the drift velocity of the electrons is zero, the magnetic field is no longer due to the movement of the electrons but to that of the atomic cores. (By the way, the change of the reference frame is minuscule: the drift velocity is a fraction of a millimeter per second.) Thus, the cause of the magnetic field changes according to the reference frame, „independently of what is happening within the system“.

Consider finally the magnetic field of an electron beam: in the reference frame of the electrons the magnetic field is zero. Thus, there is a magnetic field or there is non, „independently of what is happening within the system“.

All these are situations which are familiar to the physicist, but to which he had to get acquainted. The momentum flow in a spring is just one more example.

Disposal:

1. As far as school is concerned: Follow the Karlsruhe Physics Course. Choose the positive *x* direction, once and forever, to the right, seen from the pupils’ side. From our experience with a very great number of students, they have no difficulty with the rule that in a stretched spring momentum flows to the left.

2. At a later time the problem is discussed in another context. Here the students learn: „A change of the reference frame does not change the world, but only the description of the world.“

Friedrich Herrmann

[1] *M. Bartelmann, F. Bühler, S. Großmann, W. Herzog, J. Hüfner, R. Lehn, R. Löhken, K. Meier, D. Meschede, P. Reineker, M. Tolan, J. Wambach und W. Weber:* Gutachten über den Karlsruher Physikkurs; in Auftrag gegeben von der Deutschen Physikalischen Gesellschaft.
http://www.dpg-physik.de/veroeffentlichung/stellungnahmen_gutachter/Stellungnahme_KPK.pdf

“With such a definition a problem is created, since the direction of the x -axis can be arbitrarily fixed – and changed - in space, regardless of the physical events within the system. This means that the direction of the KPC momentum current can be arbitrarily changed, i.e. independently of events in the system, only by a new choice of the coordinate system. Hence, we conclude that the direction of the KPC momentum current is not a property of the system.”

5.25 The point in mechanics

Subject:

In the domain of point mechanics the concepts mass point, position, trajectory, force field… are employed.

Point mechanics is the favorite mechanics of the physicists. Physics students learn it in great detail, and also in physics textbooks for the school mass points are usually brought up.

Deficiencies:

1. In physics, point mechanics is so dominant, that everybody finds it natural to speak of a mass point or a point mass instead of a body. This corresponds to the idea, that there are force fields, and thus, forces that can have different values in every point of space.

In such a description of nature important concepts of mechanics loose their meaning or become problematic, such as pressure, or the densities of mass, electric charge and energy. However, often, and in particular in school physics, there is not a necessity to introduce this singular theoretical description.

2. In addition, there is the somewhat easy-going handling of the designation „mass“ and „point“. Let us briefly explain the two concepts. Mass is a physical quantity, that measures a particular property of a body or a particle: its inertia and its gravity. And what is a point? Among the many meanings (48 different meanings in the Wiktionary) the only one that is pertinent in our case is: „a specific location or place, seen as a spacial position“.

Now, the designations mass point and point mass are not consistent for different reasons.

According to our language habits a mass point would be a point that has a mass, whereas a point mass is a mass that is point-shaped. Actually, both statements are senseless.

Let us begin with the mass point: An object, a body or a particle has mass. A point, i.e. a geometrical object, cannot have a mass in principle. That does not mean that its mass is 0 kg. Rather is does not have the property that is measured by mass.

And the point mass? A body can by point-like, i.e. it can be sufficiently small. Mass however is a variable in the sense of mathematics. As such it can neither be point-like nor not point-like.

How do the textbooks explain these concepts and how do they motivate the designations? We look at two examples from university textbooks.

In one of them mass points are defined as points that possess a mass.

Our second book does it somewhat better: „The moving objects have to be idealized, … We shall call such objects point masses.“ It is not really nice to call an object „mass“, but at least it is explained that a new meaning is given to the word: Here, mass is not the name of a physical quantity but of an object.

Such conceptual looseness may have no harmful consequences in the realms of research and engineering. In school, however, conceptual carefulness is not pedantry, but it is the condition to obtain clearness in the minds of our pupils or students. Every teacher knows: When using clear concepts physics does not become more difficult but it becomes more simple.

3. Also the fact that we never speak about momentum points, entropy points or energy points might give us food for thought. We may conclude that in the minds of many physicists mass is more than a variable that describes a property of a body or a particle. Regarding electric charge, it is treated like the mass. In the minds of physicists there are not only point masses but also point charges – with the same harmful side effects: in the minds of the students the electron degenerates to a point charge. Also in this case we can observe the unfortunate confusion between the physical system or object (the electron) and the physical quantity (the electric charge).

Origin:

Regarding the dominant role of point mechanics in physics:

> The great successes of point mechanics in astronomy and its important role in particle physics.

Regarding the inconvenient wording:

> The word mass is misunderstood as synonymous to the word matter.

Regarding the point mass in the school curriculum

> The education of the teachers: one semester point mechanics in the frame of the experimental physics track, one semester point mechanics in the theoretical physics track, zero semester mechanics of continuous media.

Disposal:

1. Avoid the designations mass point, point mass and point charge. If one really believes to need the pointlikeness of a body, speak of pointlike bodies. It would be better however, to call them small bodies. Or if one has good reasons to believe that the word is not misunderstood, call them particles.

2. There is hardly any reason to introduce point mechanics at school. The mechanics of continuous media is more appropriate for applications of physics to the every-day world. Some problems with the concept of force that usually arise will then simply disappear.

Friedrich Herrmann

5.26 Newton's Third Law of Motion

Subject:

„When one body exerts a force on a second body, the second body simultaneously exerts a force equal in magnitude and opposite in direction on the first body."

„The third law states that all forces between two objects exist in equal magnitude and opposite direction: if one object *A* exerts a force $\mathbf{F}_A$ on a second object *B*, then *B* simultaneously exerts a force $\mathbf{F}_B$ on *A*, and the two forces are equal in magnitude and opposite in direction: $\mathbf{F}_A = -\mathbf{F}_B$."

„The third law means that all forces are interactions between different bodies, or different regions within one body, and thus that there is no such thing as a force that is not accompanied by an equal and opposite force."

„Forces always come in pairs – equal and opposite action-reaction force pairs."

The validity of the law is often shown in an experiment: Two persons standing on skate boards pull by means of a rope one towards the other; in one case one of them is pulling, in the other case the other, Fig. 1.

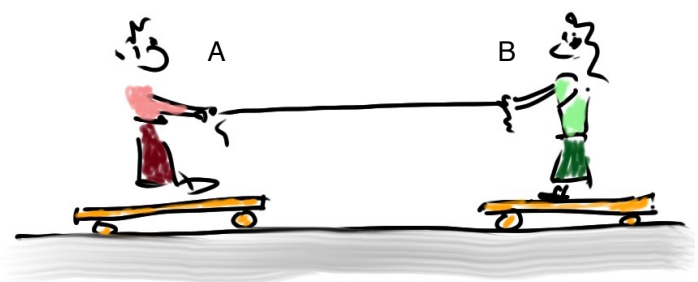


Fig. 1. Experimental proof of Newton's Third Law: One person is pulling, the other not.

Deficiencies:

I confess, that for a long time I didn't understand the law, even at the end of my physics studies. But I also confess, that I was not really interested in it. It seemed to me that it was similar to those rules or claims, that one also learns in religious instruction. You learn what you have to answer in a specific situation (namely in the examination). And of course, I knew what to say. It is not difficult to repeat the statement.

Here is my problem: From a law I expect, that it tells me how things behave, but also how they do not behave. To be understandable, I must be able to imagine a world in which the law is not valid. Take as an example Newton's Second Law (or what we call so today): It reads

$$F = m \cdot a.$$

Those who do not know it yet, might imagine that the relation between force and mass is different, for example like this:

$$F = k \cdot m^2 \cdot a.$$

However, I simply could not imagine a world in which the Third Law is not valid. How would it look like, when the force that boy A exerts on body B is not equal to that, which exerts B on A? That cannot be – for reasons of symmetry. So, why do I need a law? Since everybody who is confronted with the situation of Fig. 2 will have this uneasiness, somebody had the questionable idea of the experiment of Fig. 1.

Fig. 2. The situation is symmetrical.

At the beginning, the setup is symmetrical – two chariots, two persons –, but then the symmetry is broken by letting pull only one of the protagonists. The accompaniment is as follows: A is pulling, i.e. A exerts a force, B is not pulling. So, B does not exert a force on A? Actually B exerts a force, perhaps not intentionally. But otherwise A would not be accelerated. So one possibly does not notice immediately that the "pulling" of one or the other person has nothing to do with the Third Law. It only reveals that the experimenter is confusing statements about energy and momentum. What makes the difference between the two partial experiments – A pulls or B pulls – is only the energy source for the acceleration. The reason why one may take the experiment for convincing is that one believes in the doubtful saying that forces can be recognized by a "muscular sensation". In some proposals for the experiment that person who is not pulling does not hold the string in her hand, but has it tied around the waist, so that there is no "muscular sensation" in her arms. (It seems that the experimenter forgot that there is also a muscular sensation in the hip and the legs.)

Origin:

From Newton himself. There is no doubt that Newton was ingenious. However, at his times, it was normal for a scientist to describe the world like a mathematical object, i.e. axiomatically. The title of his work is „Mathematical principles of natural philosophy", and the text is full of *definitiones*, *leges*, *scholia*, *corollaria*, *lemmata* etc. Of course Newton missed this goal, as Ernst Mach shows it in detail [1]. So it is not a surprise that the rule of the two forces that are equal and opposite appears as one of his laws.

Disposal:

1. The third law is a (trivial) consequence of the conservation of momentum. Since the law of momentum conservation is treated anyway, no additional (third) law is necessary.

2. The experiment with the skate boards can be useful, if in addition to the momentum balance also the energy balance is discussed, i.e. if one asks for momentum and energy currents.

Friedrich Herrmann

[1] E. Mach: The Science of Mechanics, a Critical and Historical Account of its Development, The open court publishing Co. 1919, Chicago, London, p. 187

5.27 The falling cat

Subject:

You know the story with the cat. It is arbitrarily oriented thrown into the air or simply dropped: it always makes a gentle landing on its four legs outstretched. If you have some science education, you may fear that the law of angular momentum conservation will be overridden for a moment by the cat. However if you look at Wikipedia under “Falling cat problem”, you learn that everything goes right: “The solution of the problem, originally due to Kane & Scher (1969), models the cat as a pair of cylinders (the front and back halves of the cat) capable of changing their relative orientations. Montgomery (1993) later described the Kane–Scher model in terms of a connection in the configuration space that encapsulates the relative motions of the two parts of the cat permitted by the physics. Framed in this way, the dynamics of the falling cat problem is a prototypical example of a nonholonomic system (Batterman 2003), the study of which is among the central preoccupations of control theory. …

In the language of physics, Montgomery's connection is a certain Yang-Mills field on the configuration space, and is a special case of a more general approach to the dynamics of deformable bodies as represented by gauge fields…”.

Deficiencies:

I could not find any hint that the Wikipedia entry is meant as satire. Normally, it would then be eliminated after some time.

First of all, briefly what the problem is. It seems apparently surprising, if not contradictory, that the cat makes the turn. One has the feeling that there is a problem with the law of angular momentum conservation. Apparently this concern is confirmed when one reads explanations like the one quoted above. Anyway, the trick employed by the cat seems not to be a simple one.

And now the shortcomings:

1. The rotation is not a special skill of cats. Humans and other, reasonably mobile animals can do it too. Try it yourself:

Stand on smooth ground on one leg (preferably with smooth soles, or even better in socks).

Make a quarter turn around the vertical axis.

I do not explain how you have to do it, because I want to prove that you can do it without any instructions.

(You can also realize a rotation in a different way, taking advantage of the friction. Try also this standing on one leg. But that’s not our topic here.)

2. What the cat does (or what you just did) is not more remarkable than many other accomplishments that we are constantly doing, and that we are not worried about in physics lessons (perhaps wrongly): walk, run, biking, freehand cycling, ice skating, rope dancing ...

3. An unnecessary effort is made to resolve the apparent contradiction.

Origin:

1. The discussion of the problem has a long tradition. Already Maxwell and Stokes, but also many others, have dealt with it.

2. It reveals the child in man (or woman).

3. We, the physicists, can thus show to the rest of humanity, i.e. those 80% of the population, who are proud of their physical illiteracy, that physics is not only concerned with Higgs particles, entangled photons, and dark energy, for which they are not interested. Even understanding your beloved pet requires physics.

4. Maybe also a slightly inhibited relation to the angular momentum.

Disposal:

A nice effect that can be shown in class. As I said, you do not need a cat. In order to visibly exclude friction during rotation, I ask a student to sit on a swivel chair and turn horizontally without touching the floor. I have never experienced that someone could not.

What is interesting about the experiment? First of all, the fact that the analogous experiment of translation mechanics does not work.

It would look like this: Two chariots A and B; Willy (the protagonist of the Karlsruhe Physics Course) sits on chariot A and tries to pull or push against chariot B, or shake it back and forth, to shift the center of gravity of the entire system, Fig. 1.

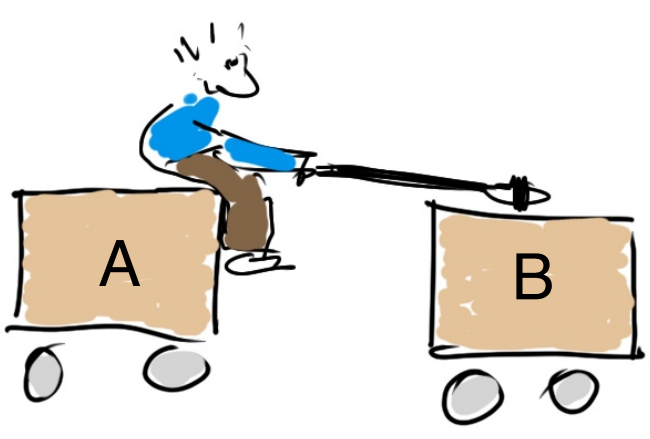


Fig. 1. the center of mass cannot be shifted

With

$$p = m \cdot v$$

and

$$\Delta s = \int v\, dt$$

we obtain

$$\Delta s = \frac{1}{m} \int p\, dt$$

and

$$m \Delta s = \int p\, dt$$

This holds for chariot A (with Willy) as well as for chariot B. Because of momentum conservation we have:

$$m_A\, \Delta s_A = - m_B\, \Delta s_B$$

To any displacement Δs_A of chariot A corresponds a displacement

$$\Delta s_B = -\frac{m_A}{m_B} \Delta s_A \qquad (1)$$

of chariot B. This means that the center of mass of the system of both chariots does not move, whatever Willy does. And if at the end the distance between the chariots is the same as at the beginning, the position of each chariot will be the same as that at the beginning. A prerequisite for this conclusion, however, was that the masses m_A and m_B are not changed.

But let us also investigate what happens when mass changes are allowed.

Chariot A was initially empty and light, chariot B loaded with sand and heavy. Willy sits again on chariot A and pulls with the help of the pole on the heavy chariot B. B moves only a little, A much. Next B is discharged and A is loaded, i.e. now A is heavy and B is light. Willy pushes himself away from B with the help of the pole. Now A moves a little and B a lot. At the end, the distance between A and B is back to the beginning, but the whole system has shifted to the right. This was possible because of changing the masses: the ratio m_A/m_B was not the same when Willy pushed chariot B away and when he pulled it back. That the center of mass of the two chariots has changed is not a surprise. We have cheated, so to speak. However, the story is interesting because we are doing a very similar thing in the rotational analogue. In this case however, without cheating.

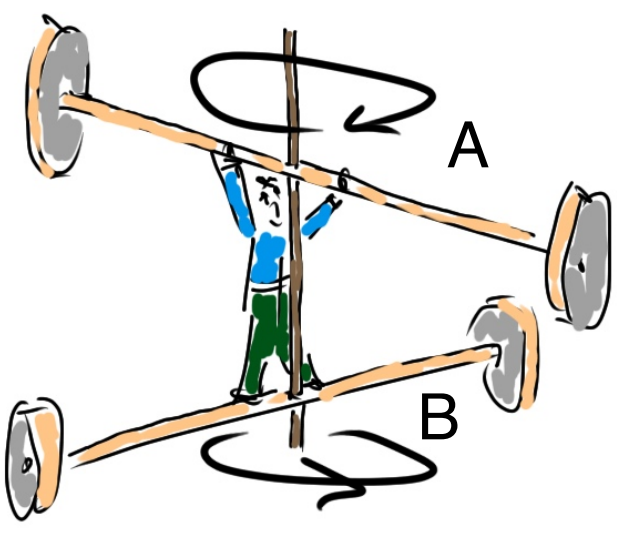


Fig. 2. By turning the dumbbells back and forth, the orientation can be changed by changing the moment of inertia of a dumbbell in the process by shifting the weights at the ends.

We consider two dumbbells that can rotate aound a common axis, Fig. 2. Instead of a net shift, we want to achieve a net rotation.

The equation analogous to (1) applies:

$$\Delta\alpha_B = -\frac{J_A}{J_B} \Delta\alpha_A \qquad (2)$$

(α is the angle of rotation, J is the moment of inertia, the derivation is exactly like the one above).

If one stands on the lower dumbbell and tries to twist the upper one, then also the lower one rotates, and the angles of rotation stand in the relation given by equation (2).

If we leave the moment of inertia unchanged, then to each $\Delta\alpha_A$ belongs a specific $\Delta\alpha_B$ given by equation (2). If we rotate once back and forth, at the end each of the dumbbells points in the original direction.

But we can change the moment of inertia without changing the mass, without loading or unloading anything. If we make a back and forth rotation, and make sure that the moment of inertia J_A during the forward rotation is greater than in the reverse rotation, a net rotation angle remains. That is what cats or humans do when they turn.

Friedrich Herrmann

5.28 Newton's Third Law (for the third time)

Subject:

„If Willy and Lilly pull with the same force, the rope does not move. There is force equilibrium. If the rope moves, say to the left, Willy pulls with the greater force."

Fig. 1. Does Willy pull with the grater force?

Deficiencies:

I am not angry with the reader of this column if he clicks this article away. After all, the third law was already twice the topic - for the last time only a few months ago. Here it is again, for a current reason.

Every author wishes many readers. But not only that; he wishes specific readers. Also the writer of this column has such a hope: this column may be read by textbook authors. Unfortunately it is not. Thus, readers who are not textbook authors can only watch with desperation or mischievousness as the schoolbook authors pass the same mistakes from one generation to the next.

Nevertheless, here is a correction note to a textbook that is just fresh from the press, see the above, slightly alienated quote: Since the mass of the rope can be neglected compared to the other participants, the amount of the force that Willy exerts on the rope is always equal to that exerted by Lilly. This is Newton's third law. Maybe it would have been a good idea to check it by measuring.

When we apply the claim to another system, there is an interesting conclusion: two electrically charged bodies attract each other, they make „tug-of-war". The „rope" in this case is the electric field. If it were to follow the rules that underlie our citation, then one of the bodies could pull with a greater force than the other. This would set the center of mass of the whole system in motion. In times of scarce energy, maybe an interesting business model – if it worked.

Origin:

1. Mechanics is difficult when formulated in the Newtonian way of speaking.

2. The claim that we feel a force by our muscular effort.

3. If a misinterpretation, explanation or other statement has no adverse consequences, then the correct interpretation has a poor chance of survival. We know that from biological evolution. The protein building blocks of all living organisms are left-handed, although right-handed molecules would not have any evolutionary disadvantage. One species happened to be in the majority, and from then on the other species had fewer and fewer chances to survive until they became extinct. Apparently, the failure to understand the third law has no adverse consequences, neither in an examination at school or university nor in everyday life.

Disposal:

If one chooses to stick with the late Baroque Newtonian way of speaking (and does not use the momentum current representation in which the difficulties do not occur), then nothing remains but to really understand Newtonian mechanics, which apparently not everyone succeeds.

If one wants to discuss the tug of war in school, here are some suggestions.

The intention is to determine whether Willy is stronger than Lilly or vice versa, where we mean by "stronger" not necessarily a greater force. The question is, first of all, what makes Willy and Lilly different in this context?

One could do the following (thought) experiment. Measure the strength of Willy and Lilly separately – with the arrangement of Fig. 2.

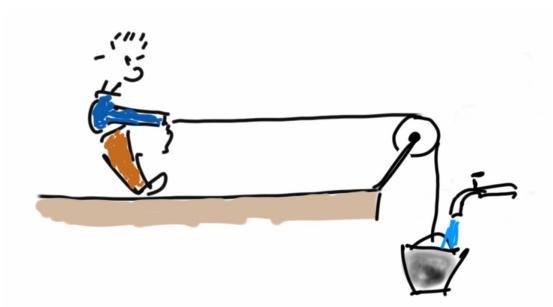


Fig. 2. Measuring Willy's strength

First, Willy must use the rope to hold the bucket in suspense. As long as there is not much water in it, that's no problem. But now water is constantly flowing, until the bucket finally becomes so heavy that Willy can no longer hold it. The amount of water is then a measure of Willy's strength. Thus, the force at which Willy can just hold the bucket is a measure of his strength.

Then Lilly's strength is measured in the same way and we decide who is „stronger" and we know who would win the tug-of-war contest. But the question remains which property or ability of the two persons, expressed in physical terms, was measured here. It could be that Willy has smoother shoe soles, and therefore loses. Of course that's not what we wanted to measure. So let's say the contact with the ground is perfect, i.e. no slipping and rubbing. Now we might see what really matters. The person, say Willy, exerts forces, on the rope and on the Earth, which are equal in magnitude. For the amount of these forces there is a maximum value that can not be exceeded because Willy collapses or falls over. Which of his muscles are failing depends on which posture he has taken.

Therefore what is measured is this maximum force. In tug-of-war, this value is reached first by one of the two partners. He or she loses the game.

Expressed in terms of momentum currents: Willy's (or Lilly's) body can only endure a momentum current of a certain strength; at a higher value, the momentum conductor breaks down, comparable to, for example, a fuse that only withstands a certain maximum electrical current and interrupts the circuit when the current becomes too strong.

Now the tug-of-war problem has another aspect. The bodies of Willy or Lilly does not just have to withstand (transmit) the force. It must first be ensured that the forces arise at all. And this is also where the muscles are needed. This time, however, not in their role as a force transmitter (momentum conductor), but as a mechanical energy source. This is needed, even if in equilibrium no energy is flowing.

It can be seen that the physical explanation of tug-of-war is more complicated than one might have guessed. It is inappropriate to use it to explain the third law, since the question of actio and reactio is rather secondary compared to the other problems.

However, here yet another suggestion: If we give a wrong explanation that has no consequences for the students, who may still become good event managers, auditors or even engineers, we could also think of omitting the topic altogether. One would gain time for something more sensible. It would not hurt the reputation of physics either.

In the case of tug-of-war, a statement that can easily be parroted, and that is not entirely implausible, but whose correctness is difficult to establish was included in the curriculum of general education. We know that also from elsewhere. Physics actually deserves a better reputation.

Friedrich Herrman

5.29 The definition of the force

Subject:

"It [the force] is defined as the rate of change of the momentum, so that for its magnitude holds:

$F = dp/dt$."

"The equation [$F = m \cdot a$] is a definition of the force that makes the external action on a body measurable through the acceleration of the body."

"The concept of force dates back to Isaac Newton, who in the 17th century created the foundations of mechanics with the three Newtonian laws, defining force as the temporal change of momentum."

Deficiencies:

I have to bother the readers again with Newton.

1. If the force were defined as the rate of change of the momentum, Newton's second law would not be a law but a definition. But it is a law because it makes a statement that can be tested by the experiment: measure the rate of change of the momentum of the considered body and, independently of it, the total force acting on the body.

2. If the force were defined by the equation $F = dp/dt$, the forces in each static arrangement would be equal to zero. Engineers would have a problem.

3. If one is convinced that the force is defined as dp/dt, this fact should reflect in the language one is using. It would then be allowed to say, for example, that body A exerts a momentum change on body B.

Is physics not brought into close proximity to theories like the so-called scientific socialism or psychoanalysis, by the rigor, which is exhibited in the school books, but that is not maintained? Physics does not deserve it. And is it any wonder that physics is the most unpopular school subject?

Origin:

The process of writing of a new textbook could be imagined as follows: When writing, the author is guided by existing older works. He detects inconsistencies, awkwardnesses and perhaps errors, and corrects accordingly. The physics books gradually become better and better. Yes, so it could be imagined – but it's not like that. The new books are sometimes getting better, but sometimes worse. A subject, that was already clear, may be disfigured or twisted again – clearly visible in the case of mechanics.

So some author believes that he is better than Newton, but he is wrong. Newton's language is somewhat difficult to read for us today, but no one has ever surmounted his logic (not even the great Ernst Mach or Ludwig Lange, the inventor of the inertial system).

Here, to recall Newton's definition of force:

Def. IV.

Vis impressa est actio in corpus exercita, ad mutandum ejus statum vel quiescendi vel movendi uniformiter in directum.

Or in English translation (by Jacob Philipp Wolfers 1872):

"An impressed force is an action exerted upon a body, in order to change its state, either of rest, or of moving uniformly forward in a right line."

Newton clearly does not say that the force is defined as the change in the state of motion of a body.

Disposal:

We can measure a force F_1 acting on a body by means of dp/dt, if we make sure that no other force F_2 acts on the body. This is sometimes easy, and sometimes less easy. But it does not mean that we *define* the force to be dp/dt.

It is found in such experiments that the relation

$dp/dt = F$,

is always valid, where F is the sum of all forces acting on the body. Thus, the experiment shows that momentum is a conserved quantity. Before measuring one might also have imagined to find

$dp/dt > F$,

or

$dp/dt < F$.

Then, in the first case, one would have concluded that momentum can be produced and, in the second case, that it can be destroyed.

The error in our quotations would not have been made if the quantity F had been interpreted from the outset as a momentum current. Then our intuition immediately tells us how the relation between F and dp/dt must be: Since p is a conserved quantity, the change of the momentum is equal to the total current strength of the momentum flowing into the body.

The momentum does not change when a current flows through the body only. Then the inflow is equal to the outflow. Or in the Newtonian language: two forces of the same magnitude but opposite directions act on the body; there is equilibrium of forces.

Friedrich Herrmann

5.30 The force in the table top

Subject:

In the mechanics section of our Physics textbooks, various forces are addressed: weight, downhill force, normal force, frictional force, buoyancy, and many others. One says that one body exerts a force on another body. If one does not want to mention the body that exerts the force, one also says that the force acts on a body. Sometimes it is also said that it acts via the line of action, and that it acts on the point of action. Occasionally, another wording is used: instead of saying that a body exerts a force on the Earth, it is said to exert the force on its support, for example, on a sloping plane.

Deficiencies:

One sees from these linguistic expressions that one is dealing with a difficult quantity (as opposed to all claims that every person has a natural feeling of a force). The fact that the linguistic handling of the quantity of force is so unusual shows that the concept is conceptually difficult. It also can be seen that most students of physics have not fully understood the concept.

Consider the case where a box is on the ground. (A simpler static problem can hardly be imagined.) In this case, the Earth exerts a force on the box that can be calculated according to $m_{box} \cdot g$. The Earth exerts the force? The whole Earth? Even that down in New Zealand? That's the way it has to be. It exerts it, as we said, on the box. On the whole box? Yes and no. Yes, on the whole box. But it acts, as one says, on a point of action. But how does it get from the point of action to the other points of the box? Especially when the box is empty, and the point of action is where there is only air. And how about the Earth? Does it also have a point of action? In any case, according to Newton's 3rd law, there is an opposing force to our first force, and it will probably act on the point of action of the Earth. Or not? Don't we rather say that the box exerts the force on its supporting surface? At least that's more plausible than on the whole Earth. Now, let's put the box on a table. That makes matters even more complicated. Now the box exerts a force on the table, or more precisely, on its supporting surface of the table top. The table top then passes it somehow on to the legs of the table. (But is it allowed to speak in this way in an exam? If not, how do we say it then?) And every table leg exerts a force on the Earth. Again: Only on the four supporting surface areas, or on the whole Earth, including New Zealand? And there is also the gravitational field. What role does this play? It is sometimes said that it "mediates the force". It mediates between two bodies, like the marriage broker between two people of different sex.

Of course you have noticed, dear reader, and you may reproach it to me, to play possum. Of course I did. But do not the questions, the awkward phrases I outlined, suggest themselves? Does one not have to ask such questions if one is introduced to statics, as happens in school (and also in the physics lecture at the university)?

The problem is that we always have to deal with a closed path in a static problem, or at least part of it – but that's not what we say. Instead, at best, we are talking about a few cross sections in this path, and at worst about points of action, which, like for a ring or an empty box, can be where the body is not.

So most of our box problem is not addressed at all: what about the forces in the box, in the table top, in the table legs, in the ground, in the gravitational field? We picked two or three places for which the calculation of the force is easy. But then we can hardly claim that we are talking about the world in which live our students.

One might think that nothing could be done here: class time is a scarce commodity. We have to confine ourselves to simple cases. Finally, we also treat only linear oscillators, and leave the nonlinear ones for the university. We treat the ideal gas and leave the real gas for the university, etc ... If we wanted to treat the entire force distribution in a static problem (more precisely, the distribution of mechanical stresses), we needed, one might think, Landau-Lifshitz volume 7 or something similar – so it's not for the school. Therefore, we conclude, we confine ourselves to describing the forces only in a few places or points.

In fact, the situation is different here than in the case of the linear approximation of the oscillations or the idealization of the gas. It's not that the force within the box is smaller than at the area where the box touches the ground. So it's not an approximation of what we do, but we just hide most of the phenomenon.

Origin:

We teach mechanics like in Newton's time, when there was no other way. Euler and Bernoulli were later and left hardly any trace in school physics. At the university, the future teachers will instead learn the theories of Hamilton and Lagrange, who are very elegant, but with whom they can hardly do anything in their profession.

Newtonian mechanics uses a language that does not even raise the question of the force in a tabletop, within the Earth, within the sphere that is pushed by another, or the force within the gravitational field. Forces simply act on the body, on the Earth or on the Moon, on a spring, and occasionally on a rope. And of course, neither on nor within the gravitational field – because that did yet not exist in Newton's times.

Disposal:

The baroque age force metaphor introduced by Newton is outdated. Newton invented it with the intention of not having to talk about the problems mentioned above – at that time an ingenious idea. His main problem was that he did not yet have the concept of the gravitational field. And of course he was still a long way from a mechanics of continuous media.

Today, we are in a much better position and no longer need the astute Newtonian language, after Euler and Bernoulli developed the mechanics of continuous media, and especially since Planck showed that forces can be interpreted as momentum currents. The description of our box on the table now goes like this: momentum comes from the Earth, from all points of the Earth through the gravitational field into the box – in a broad stream to all points of the box, where mass is located. Then it flows through the matter of the box to its bottom, from there into the Earth, where it spreads widely. So the circuit is closed. Of course, there are many more momentum currents flowing through the Earth, but the one just described is that part that has to do with our box.

All this can be said, even before one starts with any calculation.

Sometimes one gets the impression that the view prevails, as long as one does not calculate, one has not yet to do it with real physics. I do not agree: the essence of the process of understanding precedes the calculus.

When we say that the water evaporates in the ocean, is transported with the air to the land, condenses, falls as rain down to the Earth, accumulates in streams and rivers, and gets back into the ocean, we make important statements about the water cycle, without mentioning a current strength for any section through this water circuit, or a current density for any point of it. Why can't we deal with momentum in the same way? It would be that easy!

Friedrich Herrmann

5.31 Acting acceleration

Subject:

"Near the earth an acceleration of $g = 9{,}81 m/s^2$ acts on a stone."

"During the acceleration of a car, an acceleration of approx. 0.3*g*, acts on the occupants…."

"The direction in which the acceleration acts, also plays a role. Most damaging are 'downward' accelerations, which cause the blood to shoot into the brain and into the eyes."

"Observers in freely falling frames who plunge through the hole's horizon see no real particles outside the horizon, only virtual ones. Observers in accelerated frames who, by their acceleration, remain always above the horizon see a plethora of real particles." (With "Hole" a black hole is meant.)

Deficiencies:

The quantity that we abbreviate in kinematics with *a*, is called acceleration. The name fits reasonably well, because it measures what one would call acceleration in colloquial speech (apart from the fact that one assigns in this way an acceleration also a circular motion).

If one were to formulate a physically correct proposition, that is, to add to the subject of acceleration a predicate and an object, one would say that a body has an acceleration, or its acceleration has one or the other value, as one also says it has a certain velocity, temperature or density. By no means would one say that the acceleration acts on the body.

The acceleration is a kinematic quantity of the entity that moves – this entity must not be a body at all; it can also be a point on the screen of the computer – just as the velocity, about which we also do not say that it acts on a body.

Only something else or someone else can act on the body. The acceleration can be at most the effect of something.

An "acting" of the acceleration is encountered most frequently in the context of the gravitational field of the earth. Correspondingly, one calls the physical quantity *g* gravitational acceleration.

In order to avoid this awkward manner of speaking, in the German school book literature one prefers to use for *g* the term "local factor".

That is a little better, but it's awkward for another reason.

The equation

$$F = m \cdot g$$

is the analogue to the equation known from electrostatics

$$F = Q \cdot E$$

With the same argument as for *g* one could also characterize *E* as a local factor, for the value of the electric field strength also depends on the location, as well as the values of innumerable other physical quantities.

Why *g* is not called gravitational field strength, as *E* is called electric field strength and *H* magnetic field strength?

One might argue that it is pedantic to criticize such language habits. Doesn't everybody know what is meant? This would be acceptable, if it were an isolated case. Unfortunately, however, it is one of many examples that in physics one uses an unclear, inappropriate or contradictory wording. How much could physics win by a clear, coherent language!

Origin:

That *g* is not called or interpreted as gravitational field strength is probably due to the fact that one still sticks to the idea of an action-at-a-distance, reluctantly introduced by Newton. After all, at Newton's time there was no gravitational field yet.

Even if this way of speaking has no serious consequences, it is nevertheless an indication of an antiquated world view.

I also have to admit that whoever speaks like that is in good company. By way of exception, I am quoting the author of one of the above quotes, namely the last one: The sentence was written by Kip Thorne (Physics Nobel Prize in 2017, which he certainly deserves). Thus, speaking jargon of physics (and of Nobel laureates) can give one the comfortable feeling to belong to the insiders, no matter whether one understood the physics or not.

Disposal:

Never let an acceleration act on a body.

If you really want to make something acting (but it would be even better not to let anything act at all), then let the force act, or if necessary the field, or the earth, but for God's sake not the acceleration.

And call *g* the gravitational field strength. So it becomes clear that the two equations $F = m \cdot g$ and $F = Q \cdot E$ have something in common.

Friedrich Herrmann

5.32 Movement with constant velocity

Subject:

One of the simplest and at the same time most noticeable movements is that of a vehicle that travels uniformly: A car on a country road or motorway, or a train on a free track.

Deficiencies:

How does physics-teaching deal with these processes? They are mentioned and discussed in kinematics: as an example of a movement at constant velocity, the simplest movement ever.

But what do dynamics say about a regular car or train journey?

For example the following:

> "For a vehicle to move uniformly, energy must be continuously supplied to its engine. This is because friction constantly releases heat into the environment during the movement.
>
> The driving force for the uniform movement is just as great as the total frictional force F_F. The energy required to move the vehicle uniformly along the distance s is then $E = F_F \cdot s$. The kinetic energy of the vehicle remains constant during this process."

These sentences may be correct. It would be nice, however, if one would also have learned what is to be understood by the driving force. A force is always exerted by a body A on a body B. What is in our case body A and what is B? After all, the car is driven. So then the car would be body B. Wouldn't it? Now the drive comes somehow from the engine according to general speech and expectation. But the engine is part of body B, on which the driving force is supposed to act. The learner has no choice but to learn the sentences by heart and, if necessary, recite them.

The sentences also do not answer a question that the naive reader might have: Why does the velocity of the vehicle remain constant? Is the answer too difficult? Or is it trivial? Why are the two forces the same? Does the driver have to use the accelerator pedal to find the exact position where the car neither accelerates nor decelerates?

Let's ask another book what it has to say about the matter. Here the subject is addressed after friction has been discussed in all its details, with sliding, static and rolling friction, with the corresponding laws of force and with the interpretation on the molecular level. All this seems to be necessary for the understanding of the car driving uniformly.

Even if one does not understand everything, here one definitely learns: The matter is extraordinarily complicated. In order to understand the movement of the car, one has to distinguish between 10 different forces, namely driving force, driving resistance force, interaction force, static friction force, dynamic friction force, normal force, rolling friction force, air resistance force, input force and acceleration resistance force. The well-meaning reader, however, also asks himself what is meant by the driving force. The text says:

> "The driving force F_A, which is transmitted from the engine via the gearbox to the wheels, can at most be equal to the maximum static friction force."

So here it is clearly stated: The driving force comes from the engine. Let us try to understand. Let us assume that the pistons of the engine move in a vertical direction, i.e. up and down. The hot gas presses on the pistons. Of course it also pushes downwards and to the sides, but that probably doesn't matter, because the engine is driven by the moving piston. So we have a force of the gas upwards. However, the car should not move upwards, but forward. What now? It's really a problem, because even the engine as a whole can't generate a forward force. Apart from the problem we already had before: The engine is part of the car after all. So it should not exert any forward force at all, because then itself would have to move backwards.

Stupid remarks? Perhaps. But could it also be that the author has got caught in the jungle of forces he has created and that he has mistaken force for energy? Because with energy the sentence becomes correct: it goes from the engine via the gearbox to the wheels, or "is transmitted" if you want to express it more scholarly.

Origin:

With Galileo's discovery of the law of inertia, with the whole work of Newton, the writings of Descartes and Huygens, a new dawn of science began, a continuation of something that had begun about 2000 years ago in Greece, but soon fell into a twilight sleep lasting many centuries. Ever since Galileo and Newton, we know that forces cause accelerations. This insight was great. However, it also had a negative side effect: the friction, which in retrospect had led to Aristotle's rather unfortunate interpretation of the movement of bodies, now appeared only as a disturbance of the beauty of the new building of science. The true physics, it now seemed, took place in a frictionless universe. The horse-drawn carriage at Newton's time or the high-speed train today, fight only against this disturbance while they are driving. Proper physics can at best be observed during the short phase of acceleration at the beginning of the movement.

So Newton's second law became the sanctum of physics, even if in the light of the discoveries and insights that followed, it turned out to be no more than the expression of the conservation of a physical quantity, namely momentum. The law of momentum conservation is indeed an important physical law, but it is not necessarily more important than the laws of conservation or non-conservation of other extensive quantities, such as energy, electric charge, entropy or angular momentum, which were subsequently discovered.

Yet another remark about the many forces: Newton's ideas are now more than 300 years old, and much has happened since then. Thank God we no longer need the force metaphor, as ingenious as it was in Newton's time. If one uses the fact that forces can be interpreted as momentum currents, then one discovers that several of the above mentioned forces are simply one and the same momentum current measured at different locations or with differently oriented surfaces.

Disposal:

Movement at a constant velocity in the presence of friction is a beautiful subject for school – it is important, not trivial, but also not too difficult. It is a simple example of what physics calls a steady state: the outflow adjusts itself in such a way that it is equal to the inflow.

This applies to the water that flows into a container with a hole, Fig. 1. The container is initially empty. One opens the tap and lets the water run. The water level rises; this causes the outflowing water current to increase. It increases until the outflow is equal to the inflow.

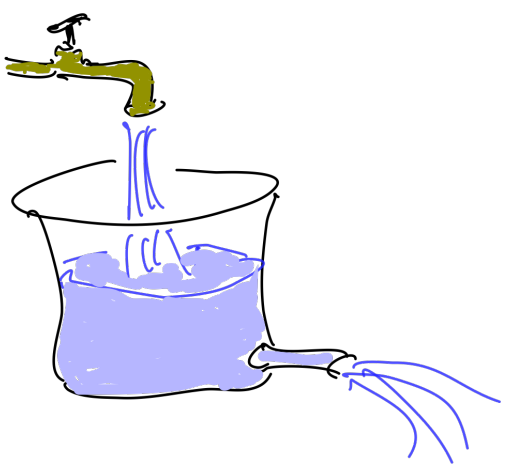


Fig. 1 The water level adjusts itself so that exactly as much water flows off as flows in.

The same holds for the car. The motor causes a momentum current to flow from the earth into the car. The velocity of the car increases. Thereby the the outflowing momentum current (due to friction) increases. It increases until the outflow is equal to the inflow.

We can say the analog about a room that is heated. First the temperature increases. Thereby... etc.

The mean temperature of the earth's surface is another example. It also is the result of the establishment of a steady state.

One could argue, that in the excerpts quoted above, much more is said than just the establishment of a steady state. Let us assume that the texts did not contain any errors: even then they are inappropriate. The need for action in physics teaching today is not so much which new topics could be introduced. Before we include new topics, we have to decide what we throw out instead. Among the ten forces related to the car there would be good candidates.

Friedrich Herrmann

5.33 Gravitational acceleration

Subject:

The factor $\vec{g}$ in the equation

$$\vec{F} = m \cdot \vec{g} \qquad (1)$$

is usually called *gravitational acceleration* or *free fall acceleration*.

Deficiencies:

1. We first write the equation for the vertical components of the vectors $\vec{F}$ and $\vec{g}$ and interchange the order of the factors in the product:

$$F_G = g \cdot m. \qquad (2)$$

So the equation tells us that the gravitational force F_G is proportional to the (gravitational) mass. If the mass is given, we can calculate the gravitational force. g is the factor of proportionality.

For a freely falling body, i.e. a body on which only the gravitational force F_G is acting, the following applies

$$F_G = m \cdot a$$

where m is the inertial mass. With equation (2) follows:

$$a = g\ .$$

Thus the acceleration of the body during the process of falling is equal to the factor of proportionality g in equation (2). Hence, g has inherited the name acceleration from a. To distinguish it from other accelerations, g is called gravitational or free fall acceleration. However, equation (2) also applies when nothing is accelerated, or when, as in the case of falling with friction, the acceleration has a different value than g. Should g really be called acceleration, just because its value corresponds to the acceleration in a special process, namely in free fall? Probably not.

2. Equation (1) can also be read differently: as a definition of the vector quantity $\vec{g}$. One determines $\vec{g}$ from the easily measurable quantities $\vec{F}$ and m. The body now only serves to measure a property of the surroundings of the earth: We measure m and $\vec{F}$, divide $\vec{F}$ by m, and obtain $\vec{g}$. If we do the same with another body, which has a different mass, we get the same value of $\vec{g}$. So $\vec{g}$ describes something, which has nothing to do with the body. But with what does it have something to do? It characterizes a property of an (invisible) entity, which is located in the surroundings of the earth, and which we call gravitational field. Accordingly, the quantity should be given a name that refers to this entity: *gravitational field strength*.

Doesn't that sound familiar to us? Of course it does. It is the way we deal with the equation

$$\vec{F} = Q \cdot \vec{E} \qquad (3)$$

We use it to measure the distribution of the quantity $\vec{E}$ in space – a quantity that describes something that has nothing to do with the test charge Q. $\vec{F}$ describes the electric field that is present even if there is no test charge anywhere.

3. $\vec{g}$ can be transformed away or it can emerge by a change of the reference frame. At a place close to the surface of the earth in a reference frame at rest relative to the earth, g is equal to 9.8 N/kg. In a freely falling reference frame at the same location we have $g = 0$.

g has this property in common with many other physical quantities: with velocity, momentum and kinetic energy, but also with the electric and magnetic field strengths. However, the insight into this fact in the case of the gravitational field is harmed by the introduction of a whole series of additional designations: Multiplied by m, this results in "fictitious forces", "g-forces", "inertial forces". As a result, one loses sight of the fact that one is always dealing with one and the same physical quantity, which, like many other quantities, assumes a different value depending on the reference frame.

4. Regarding the unit of measurement. Of course, the unit m/s^2 is correct. However, it is not a good choice because it suggests to interpret g as an acceleration. An alternative would be N/kg. This follows from equation (1). However, this unit is not much better because the field strength has only sometimes something to do with a force: only when a test body is placed into the field. Let us therefore try to orient ourselves on electromagnetism. What about the unit of measurement of the electric field strength? As is well known, mostly V/m is used, but N/C could do as well. Again, one unit of measurement is not more justified or more appropriate than the other. But what kind of unit would we like to have? If the quantity describes an intrinsic property of the field, shouldn't it merit a name of its own? Certainly; but it didn't get it. But let's have one more quick look at magnetism. What a surprise! There are two different proper names for the units that describe the field: the *Gauss* and the *Tesla*. We see again: The development of physics was and is somewhat erratic.

Origin:

Newton was around 1700, when no fields existed in physics (even though they would have perfectly fitted into Newton's mechanics). So a way of dealing with gravity (as with the whole of mechanics) was created or invented at that time, which operates with actions at a distance, and which does not contain any reference to the properties of what takes place between the gravitating bodies. Of course, under these conditions g could not express a property of something that is located between the bodies. It could only be interpreted as a factor that refers to a body. So, without a body there is no g. At that time the term acceleration due to gravity was a natural choice.

Disposal:

Introduce g via

$$\vec{F} = m \cdot \vec{g}$$

just as you would introduce $\vec{E}$ via

$$\vec{F} = Q \cdot \vec{E}$$

and call it gravitational field strength.

It turns out that the acceleration of a freely falling body is equal to this field strength. This is due to the equality of gravitational and inertial mass, which in the context of classical mechanics does not appear as an identity, but as a surprising, almost unbelievable but nevertheless to be accepted result of the observation. This should be addressed.

Friedrich Herrmann

5.34 Potential energy (add-on)

Subject:

For the potential energy we found the following statements or definitions in various text books and other sources:

1. (University) “If a body of mass m is lifted near the ground to the height h, a work
   $W = E = mgh$
   is performed against the force of gravity mg.
   It is contained in the body as energy; one can transform it at any time into the same amount of kinetic energy by letting the body fall.”
2. (School) “In order to lift a body of mass m on the earth by the distance Δr, energy must be supplied to it. Thereby its potential energy increases by the amount ... ”
3. (Wikipedia, keyword ‘Potential Energy’) “For a movement against the force of weight, work must be done on the body, which is now stored in it as potential energy.”
4. (University) “… V is called the potential energy of the mass point m ”
5. (University) “This work is stored as potential energy mgh in the system consisting of the earth and the skier".
6. (School) “The positional energy or potential energy of the system earth-body of mass m in relation to an arbitrarily chosen reference level is $E_{pot} = m \cdot g \cdot h$.
   …
   The tensional energy or potential spring energy of a spring with the spring constant D, which is elongated by the distance s from the relaxed state, is ....".
7. (University) “In many cases, the work performed on a system does not lead to a change in the kinetic energy as with a single mass point, but is ‘stored’ as so-called potential energy”.
8. (University, theoretical physics) “Besides the kinetic energy we define the potential energy V by...
   $V = E_{pot} = -\int F dx$. “
9. (Brockhaus, A German encyclopedia, 1926, keyword ‘Energy’) “The mechanical energy inherent in a body is the result either of its position with respect to the environment (e.g. in the case of a lifted load or of dammed up water) or (e.g. in the case of elastic bodies) of the position of its smallest parts with respect to each other (energy of position, potential energy)…”.
10. (University, 1936) “The ability of a body to perform work as a result of its position or the arrangement of its parts, etc., is called its potential energy. The potential energy... is measured by the product of the acting force…”
11. (Brockhaus 1910, keyword ‘Energy’) “The energy can either be actual, i.e. consist actually in work performance, kinetic energy, or it can be present without really doing work, as resting, potential or static energy”.

Deficiencies:

For the energy (as for other extensive quantities as well) one can specify a density and a current density, i.e. one can say of the energy where it is located, i.e. how it is distributed in space.

What about the potential energy in this respect? Where is the potential energy located in different situations?

This is a question to which students and pupils either get no answer at all, or one of several answers that do not quite fit together.

1. Let us first look at the case where a body on the earth is lifted.

Among our quotations, the first is the clearest: the potential energy is contained in the body. The “contained within” tells us explicitly where the energy is located.

Quotations 2 and 3 also express it clearly: the potential energy of the body increases or is stored in it, and also a textbook of theoretical physics, quotation 4, admits this: V is the potential energy of a mass point. (By the way: Would the author also say m is the mass of an energy point?)

There is a problem however. If the energy is contained in the body or object, do we have to conclude that the potential energy of the moon is contained within the moon? Or let us consider a binary star system with two stars of equal mass. Which of them has the potential energy? Or let’s come back to our initial example: a small body, such as a stone, “in the gravitational field of the Earth” (as one likes to say). One could also turn it around: One displaces the earth in the gravitational field of the stone. One gets the same value for the potential energy. Whose potential energy is it? This time that of the earth?

Therefore, some authors are more cautious. For example, in quotations 5 (university) and 6 (school), energy is stored in the system consisting of the earth and another object (a skier or a body). This would mean that in the case of the moon it is stored in the system earth-moon, and in the case of the binary star system in the binary star system. But how can this be understood: Do the two bodies share the energy in some way? If so, in what proportion? The authors do not say that they possibly have in mind that also a field belongs to the system earth-body. Do they think that the reader is not yet mature enough for the idea?

Still other authors are even more cautious: The energy is simply stored, quote 7. It is not revealed where it is stored. However, it is a reasonable expectation of the readers to learn this, because when one speaks of “storing” something, one is clearly professing a substance-like idea of the energy, and if one stores any substance or quantity, there must be a place where it is stored.

2. From the way the potential energy is introduced in university textbooks, namely as a path integral over the force in a “conservative force field”, see our quote 8, it follows that the concept is not limited to gravitational fields. Thus, one can also formulate the corresponding propositions when moving a “test charge” in the electric field of another charged body, or simply when pulling on an elastic spring. In this case it is obvious that one should not say that the energy is located within the test charge or within the hook at the moving end of the spring. There is no doubt that the energy is within the electric field or within the spring, respectively. This is what is said in our quote 6.

A formulation from an encyclopedia from 1926 (quote 9) is almost amusing. One can see how the author is reluctant to include a field in the explanation of the processes. But the quote also tells us what historical burden we still carry around with us.

3. Regardless of where one finally locates the energy, the impression is always created that the energy is essentially a mechanical quantity, which places a large stone in the way of later teaching topics.

4. Finally, the name: Why should the energy be called potential just because it is stored in a system at rest? After all, a charged capacitor is not said to have potential charge on its plates.

Origin:

It seems that several things are coming together.

Although it was clear from the beginning (when the energy was introduced by Joule and Mayer) that it is a quantity that describes both mechanical and thermal processes, the belief, widespread in the 19th century, that the world can essentially be explained mechanically has probably left its mark. Certainly Hamilton’s point mechanics contributed to this view. The works of Planck, Poynting and Heaviside, and the beautiful review article by Gustav Mie from 1898 [1] could not change anything in this respect, just as the development of thermodynamics by Gibbs and Helmholtz. Why does one still learn a theorem of conservation of mechanical energy, but not one of electrical or chemical energy? “Energy is an integral of motion” is the credo of theoretical physics. Frictional processes are an evil in this world that one tries to avoid.

And once again regarding the term “potential”: a body located at a great height above the earth’s surface has more energy than one located further down. How can this be? One should expect that one can somehow notice wether an object has much or few energy: by the fact that it moves, that it is hot, that it is under pressure... But our two bodies do not differ in anything but their “position”. Nonetheless, it is said that one of them has more energy than the other. Under these conditions, “potential” seems to be the right term to describe the situation. Indeed, as long as fields were not known, what distinguished the one body from the other was only its position. Potential energy seemed to be a kind of promise: “If you give me the opportunity, I will perform work.”

But even after the identity of mass and energy had been discovered, the old language was still used, see quotes 10 and 11. In a respected university book (quote 10), which was still printed in 1957, energy is an “ability”.

Disposal:

Say always clearly where the energy is located, by how much it changes, where it flows to. (Do the same for the other extensive quantities, especially momentum and entropy).

When pulling on a spring, the energy is stored within the spring. In a normal, Hookean spring, the energy is equally distributed over the length.

When displacing a charged body in the presence of another charged body, it is stored in (or taken from) the electric field. This means the really existing field, not only the field of one of the two bodies. The formula for the energy density should be known by every student.

When charging a battery, it is stored…etc..

One may see a problem with the gravitational field: here, the energy density is negative. This is unavoidable as long as the description is not done with general relativity. When a body is falling, its energy increases. This energy comes from the gravitational field, the shape of which is determined by the earth and the body under consideration. The energy of the field decreases, i.e. the absolute value of the negative field energy increases. To describe the energy flow, a second field, the “gravitomagnetic” field, is needed, just as in electrodynamics the magnetic field is needed in addition to the electric field to calculate the Poynting vector.

The gravitomagnetic forces (which correspond to the magnetic forces in electrodynamics) are immeasurably small for most terrestrial applications because of the smallness of the coupling constant of gravity. In contrast, the gravitomagnetic field manifests itself clearly via the energy current. Even if we do not want to calculate the current distribution, it is important to know that it can be done, because only then it is possible to use a coherent language to describe gravity: The energy that the falling body receives comes from the gravitational field. We do not need to say that it is only “potential”.

Friedrich Herrmann

5.35 Equations of motion

Subject:

What is meant by an *equation of motion*? Here are some answers from various school and university textbooks.

- $\frac{d\vec{p}}{dt} = \sum_i \vec{F}_i$
- $\vec{F} = m\ddot{\vec{r}}$
- $\frac{d\boldsymbol{P}}{dt} = \boldsymbol{F}$ and $\frac{d\boldsymbol{r}}{dt} = \boldsymbol{v}$
- $\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{q}_i} - \frac{\partial L}{\partial q_i} = 0$
- $\dot{q}_i = +\frac{\partial H}{\partial p_i}$ and $\dot{p}_i = -\frac{\partial H}{\partial q_i}$
- $\frac{d\vec{v}}{dt} + \mathrm{grad}(\vec{v})\cdot\vec{v} + \frac{1}{\rho}\mathrm{grad}(p) = \vec{k}$
- $\rho\frac{D\vec{v}}{Dt} = \rho\left(\frac{\partial \vec{v}}{\partial t} + (\vec{v}\cdot\nabla)\vec{v}\right) = -\nabla p + \mu\Delta\vec{v} + (\lambda+\mu)\nabla(\nabla\cdot\vec{v}) + \vec{f}$
- $\dot{\mathfrak{P}} = \frac{d}{dt}\{\mathfrak{u}M + M\,\mathfrak{w}\times\mathfrak{r}\} = \mathfrak{K}$

  $\dot{\mathfrak{L}} = \frac{d}{dt}(I * \mathfrak{w}) = \mathfrak{D}$
- $H|\psi\rangle = i\hbar\frac{\partial}{\partial t}|\psi\rangle$
- $s = \frac{1}{2}a\cdot t^2 + v_0\cdot t + s_0$
- $s = v_0\cdot t + s_0$
- $s = s_0\cdot\sin(\omega\cdot t + \varphi)$
- $s = s_0\cdot e^{-k\cdot t}\sin(\omega\cdot t + \varphi)$

Deficiencies:

The *equation of motion* seems to be an important concept. Sometimes it is highlighted in bold. But what does the term stand for? For the displacement-time law of a moving body? For a momentum balance? For the time evolution of a wave function?

I have never used the word myself, neither in lecture, nor in school lessons, and I admit that it was because I was afraid of saying something wrong – until I realized that you can't say anything wrong. The word almost always fits.

Probably also the author of the entry in the *Encyclopedia Brittanica* could not solve the problem when he formulated these sentences:

> **Equation of motion**, mathematical formula that describes the position, velocity, or acceleration of a body relative to a given frame of reference. Newton's second law, which states that the force *F* acting on a body is equal to the mass *m* of the body multiplied by the acceleration *a* of its centre of mass, *F* = *ma,* is the basic equation of motion in classical mechanics.

Origin:

Probably again the tendency to regard the kinematic aspect of mechanical processes as the most important feature, so that even in equations which clearly make a statement about momentum or energy, one may want to remember their kinematic cousins.

Disposal:

One will hardly agree on which of the equations deserves the name. Therefore my recommendation: get rid of it. One can easily get over the loss.

Friedrich Herrmann

5.36 Central force and centripetal force

Subject:

What is meant by a centripetal force? Here are some answers:

1. “A centripetal force is that by which bodies are drawn or impelled, or any way tend, towards a point as to a centre.” (Newton)

2. “Since we must conclude from the occurrence of every acceleration that a force is acting, we recognize that for the preservation of a curvilinear motion a force directed to the center of curvature is necessary, which we measure by ... . This force is called central force, also known as centripetal force, its magnitude C is:…”(Helmholtz)

3. “When a body of mass m performs a uniform circular motion, it must be subjected to a force of magnitude

$$F = m\frac{v^2}{r} = m\omega^2 r$$

which always points to a fixed point, the center (centripetal force).”
(University textbook)

4. “… This force is called centripetal force. Note that the centripetal force is not a new type of force. It is merely a name for the force that causes a centripetal acceleration, and thus causes a circular motion.” (University textbook)

5. “The centripetal force directed to the center of rotation, which retains a body of mass m with velocity v on a circle with the radius r, is

$$F = -m\omega^2 r\,.$$”

(Schoolbook)

6. “For this force directed to the center, the terms central force or centripetal force are also common.” (Schoolbook)

Deficiencies:

1. There are too many names for forces in physics. Most of them are superfluous. Often they are ambiguous.

2. Usually the terms centripetal force and central force are used in this way:

A central force is a force which, seen from the body on which it acts (body A), is directed towards another body B. The change of momentum of A is opposite to that of B. In general, a central force is not perpendicular to the direction of motion of A.

A centripetal force is a force or component of a force that is perpendicular to the direction of motion. As the body under consideration moves along its trajectory, the force vector does not necessarily point to a fixed point. In general, it is not possible to specify a body that exerts it and changes its momentum accordingly.

In the case of a circular motion, the centripetal force can be a central force. Therefore both concepts are often identified, see the quotations 2 and 6.

In the case of a rotating ring, one will speak of a centripetal force, but rather not of a central force.

3. Some authors use the designation centripetal force only for uniform circular movements. But then it is not really clear why the transverse component of the force should not always be called centripetal force. If a car drives on a curved track, i.e. not on a circular track, one speaks of the centrifugal forces (which occur in the reference frame of the car). Why should the counterforce not keep its name here?

4. In our quote 4 it is emphasized that the centripetal force is not a “new kind of force”. But how to decide whether two forces are of different kinds? We can help: Considering that the word force stands for momentum current strength, we can say it more clearly. There are two different kinds of force if the conductors of the momentum currents are of different nature, for example an elastic spring and an electric field or a gravitational field…

5. It is awkward to say that

$$F = -m\omega^2 r \qquad (1)$$

is the centripetal force. One better says that the formula

$$\frac{dp}{dt} = m\omega^2 r \qquad (2)$$

allows to calculate the momentum change of the considered body, because all three quantities on the right side of equations (1) and (2) are quantities that describe the state of the body. Only Newtons second law tells us that this change of momentum is caused by a force (i.e. an in- or outflux of momentum).

6. In our context, we are always concerned with a change of momentum and the associated momentum transport. This can happen either convectively or conductively. Equation (2) doesn’t say anything about the nature of the current. It only describes the change of momentum. According to our quote 5, a convective momentum transport would also be a centripetal force, for example, if the body is kept on its orbit with the help of a water jet coming from outside.

Origin:

Newton himself introduced the term centripetal force. One can understand his concern: For him the main issue was the orbit of the moon and the planets, and it was convenient for him to introduce a word of its own. One must also consider with which detailedness he spreads out on 500 pages what we call today classical mechanics.

Disposal:

The term central force is useful insofar as it allows, for example, to distinguish electric from magnetic forces.

The designation centripetal force, on the other hand, does not make anything clearer and does not simplify anything. To distinguish between the longitudinal and the transverse component of a force, we do not need a new name.

Friedrich Herrmann

5.37 Centrifugal force

Subject:

What is meant by centrifugal force? Let's ask Wikipedia (German page):

> "According to d'Alembert, this *basic equation of mechanics* is written in the form
>
> $$\vec{F}_{Zp} = m\vec{a}_{Zp}$$
>
> and formally interprets the second term as a force. This force is called centrifugal force F_{Zf}. It is an inertial force, more precisely a d'Alembert inertial force. It holds
>
> $$\vec{F}_{Zp} + \vec{F}_{Zf} = 0$$
>
> and thus
>
> $$\vec{F}_{Zf} = -\vec{F}_{Zp}$$
>
> The centrifugal force is always equal and opposite to the centripetal force."

Or let's look up in Tipler [1]:

> "If we want to apply Newton's second law **F** = *m***a** in an accelerated reference frame, we have to introduce fictitious or pseudo forces which depend on the acceleration of the reference frame. These fictitious forces are not really transmitted. They merely serve as a tool to ensure that the relation **F** = *m***a** also holds for accelerations **a** measured in non-inertial frames.... For an observer on the disk, on the other hand, the body is at rest and is not accelerated. Instead of **F** = *m***a**, this observer must introduce a fictitious force of magnitude mv^2/r acting radially outward on the body and balancing the pull of the string. This fictitious force directed outward, the so-called centrifugal force, appears quite real to the observer on the disk."

Or in an older German university textbook [2]:

> "This force is called centrifugal force. By stretching a spiral spring or a rubber thread, the observer can measure the magnitude of this force for the individual points in space. He finds that he is located within a force field, ..."

Deficiencies:

1. To every force belongs a mechanical stress – either within a material medium or within a field. Mechanical stress is a local quantity; it is distributed in space. But for the centrifugal force no mechanical stress can be specified. Now, the centrifugal force, according to the saying, is only a pseudo force. Therefore, one could think that one simply would have to introduce "pseudo stresses". But while for the centrifugal force and every other pseudo force one can still specify a value, this is no longer possible for a mechanical pseudo stress.

2. If one says about a body which moves on a circular path around a center, its momentum is zero, and there is equilibrium of forces, namely centripetal force equals centrifugal force, then one should call this momentum, which in this case has the value zero, pseudo or fictitious momentum. It seems that nobody has come up with this idea yet. It would be probably too obvious that something is distorted.

3. What is being done here can be clarified with the help of an analogy.

We use the fact that every force can be interpreted as a momentum current. Then, Newton's second law

$$\frac{d\vec{p}}{dt} = \vec{F}$$

tells us that the momentum of a body changes as a result of a momentum current flowing into or out of it.

This statement is analogous to the statement of the equation

$$\frac{dQ}{dt} = I$$

which tells us that the electric charge in a region of space changes only if an electric current flows into or out of the region. If we now proceed as in the introduction of the pseudo forces, we could say: We declare that the electric charge does not change in time, that is we have $dQ/dt = 0$. Now, the above equation is no longer correct. However, we can remedy the problem by introducing an electric pseudo current, and we trust that it does not worry anybody that it is not possible to say where this current comes from, how is its current density distribution and why it has no magnetic field.

4. Sometimes it is said that the centrifugal force defines a force field, see our third quotation. We do not know whether the author ever tried to draw field lines for a centrifugal force field. It would have sources or sinks everywhere in the empty space.

It is instructive to consider the derivation of the centrifugal or Coriolis force: First, the acceleration in the rotating reference frame is calculated. This derivation is purely kinematic-geometric. It yields what is called the centrifugal acceleration (and the Coriolis acceleration). Such a derivation is correct. One can describe the kinematics of a movement in an arbitrarily whirling reference frame, and obtain arbitrarily complicated results for velocity and acceleration. And one can give names to the accelerations, as it is usual in the case of the rotating reference frame. There is no objection to this, except that one does not do oneself any favor, if one chooses the reference frame clumsily. Now, there can be quite good reasons for choosing a rotating reference frame, for example in meteorology, which is interested in the movement of air and water, always relative to the rotating earth. Things become ugly only if one claims that the motion is caused by forces. Only then one gets the above mentioned inconsistencies.

Origin:

The centrifugal force can be found in the literature long before Newton, for example in the works of Descartes and Huygens. In his *Prinzipia* Newton also refers to it again and again.

How was it possible for such a strange concept to arise? Is it possible that Descartes, Huygens, Newton, D'Alembert and Coriolis were not scientifically up to date? Of course, they were. But the mental edifice, which they erected, could appear sustainable at that time only because they operated with actions at a distance. With Maxwell, the time of actions at a distance should have come to an end. But in the so-called classical mechanics, as it is taught at school and university today, the step has not yet been taken – actions at a distance are still omnipresent.

The Newtonian language ("body A exerts a force on body B") skillfully sweeps the question about who transfers the force or which system conducts the momentum current under the carpet. There are only two participants in the process: body A and body B. The ugliness of the centrifugal force consisted only in the fact that body A is missing. From a more modern point of view, according to which forces can be interpreted as momentum currents, the problem with the pseudo forces is greater: not only the source of the momentum current is missing, also the third participant is missing: the system through which the momentum gets from A to B. The momentum shows up in body B, emerged from nowhere.

Nevertheless, we have found a cautious criticism in a textbook for theoretical physics [3]. The author discusses a simple situation in two reference frames: once in an inertial frame and once in an accelerated frame, in which pseudo forces are used for the explanation. He shows that both descriptions are mathematically possible, but finally judges:

> "If the 2nd way often leads formally faster to the goal, one must nevertheless keep in mind that the 1st way of the consideration usually does better justice to the physical facts."

Disposal:

If possible, do not describe a process in a rotating reference frame, according to the rule: Choose the reference frame in such a way that the description is as simple as possible.

Regarding meteorology: There is nothing wrong with centrifugal and Coriolis *accelerations*.

[1] P. A. Tipler: *Physik*, Spektrum Akademischer Verlag, Heidelberg, 2003, S. 114 und116

[2] R. Tomaschek, *Grimsehls Lehrbuch der Physik*, Verlag B. G. Teubner, Leipzig, 1936, S. 65

[3] G. Joos, *Lehrbuch der Theoretischen Physik*, Akademische Verlagsgesellschaft, Frankfurt am Main, 1959, S. 110

Friedrich Herrmann

6

Relativity

6.1 The energy mass equivalence

Subject:

Einstein's energy mass relation $E = mc^2$.

Deficiencies:

In many schoolbooks and magazines we find the statement that Einstein's energy mass relation means that mass and energy are different manifestations of the same physical quantity, and energy and mass can be transformed one into the other [1]. If this statement was true, we could distinguish energy from mass. A decrease of energy would be associated with an increase of mass and vice versa. However, it is not true, and it is not what Einstein's relation tells us. According to this relation, mass and energy are the same physical quantity, measured with different units.

Origin:

Possibly the culprit is Einstein himself:

"It follows from the special theory of relativity that mass and energy are both but different manifestations of the same thing, a somewhat unfamiliar conception for the average mind. Furthermore, the equation … in which energy is put equal to mass, multiplied for the square of the velocity of light, showed that a very small amount of mass may be converted into a very large amount of energy and vice versa. The mass and energy were in fact equivalent, according to the formula mentioned above."

Instead of saying "may be converted into" he should have said "corresponds to".

Disposal:

Teaching should make clear the following:

1. The quantity known before as energy also has the properties of the quantity known before as mass, namely weight and inertia. A charged battery is heavier than an empty one. Hot water is heavier than the same amount of cold water, a moving body is heavier than the same body at rest, and so on. The weight differences in these examples are so small, however, that it is impossible to measure them.

2. The quantity known before as mass has also the properties of the quantity known before as energy. At a first glance, this assertion seems unbelievable. A typical property of energy is that it allows us to do some useful work. So one might expect that with 1 g of sand one should be able to realize a work of $E = 1\ \text{g} \cdot c^2$ 10^{14} J, what is obviously not true. However, we can never take profit of all the energy contained in a system. With "compressed" air of 1 bar we cannot drive a jackhammer, with "warm" water of ambient temperature we cannot drive a thermal engine. With gasoline alone we cannot run a motor. We also need oxygen. So it should not be surprising that we cannot run or drive anything with 1 g of sand alone. We also need 1 g of anti-sand. But if we had the anti-sand, it would work.

[1] "…This pair annihilation is the conclusive proof of the famous Einstein's law $E = mc^2$ for the transformation of mass into energy."

Friedrich Herrmann

6.2 The way of writing the equation $E = mc^2$

Subject:

The way of expressing the energy mass equivalence by means of the equation $E = mc^2$.

Deficiencies:

According to an old custom in mathematics a linear relationship between a independent variable x and a dependent variable y is written in the form

$y = ax + b$,

and not

$y = b + ax$.

If the relation is quadratic we write

$y = ax^2 + bx + c$,

rather than

$y = xb + c + x^2a$.

The convention helps us to grasp rapidly the content of the equation. The custom of writing the constant in front of the independent variable is established also in physics.

When reading the equation

$$E_{kin} = \frac{m}{2}v^2$$

we immediately understand that there is a quadratic relationship between velocity and kinetic energy. The equation suggests to think of a process in which the velocity may change, whereas the mass m is rather perceived as a constant. Otherwise we would write the relation as

$$E_{kin} = \frac{v^2}{2}m \ .$$

Similarly we write

$U = R \cdot I$ and not $U = I \cdot R$, or

$Q = C \cdot U$ and not $Q = U \cdot C$, or

$E = h \cdot f$ and not $E = f \cdot h$, or

$Q = I \cdot t$ and not $Q = t \cdot I$.

In each of these cases the quantity that is considered a variable in a process is placed on the right side. The quantity that is hold constant stands left of it.

According to this convention, the equation

$E = mc^2$

would be read: the energy is proportional to the square of the velocity of the light, the coefficient of proportionality being the mass. Actually the equation means something quite different: the greater the mass of a particle or body, the more energy it has, where the coefficient of proportionality is c^2.

From this point of view it would be more convenient to write the equation as:

$E = c^2m$.

But even in this form the expression has a flaw. Why should we write a coefficient of proportionality in such a camouflaged form?

Origin:

Einstein has written the equation in this form, and nobody has thought of changing it. One might speculate about the reasons. Perhaps because in

$$E_{kin} = \frac{m}{2}v^2$$

we also write first mass and then velocity.

Disposal:

Write the equation as you are used to write this type of equations:

$E = k \cdot m$.

Here k is a universal constant.

Friedrich Herrmann

6.3 Speed of light and speed limit

Subject:

The constant c in the equation
$E = m \cdot c^2$
is called the speed of light.

Deficiencies:

The Theory of Special Relativity can be derived from Newtonian mechanics by adding one new axiom: the energy mass equivalence. When doing so, it turns out that there is a universal speed limit. When momentum is supplied to a particle (or a body), it approaches this limiting velocity c. The smaller the rest mass of the particle, the faster the velocity converges to this speed limit. If the rest mass is zero the particle can move only with the limiting velocity. The value of this velocity can only be found experimentally.

When calling this limiting velocity “speed of light” the impression results that light plays a special role in the Theory of Relativity. It appears that all the other particles have to comply with the light. We believe that this is not a fortunate view of things, since all particles independently obey the same laws. There is nothing special about light, except that its rest mass is zero. But even in this respect it is not unique.

Origin:

Usually, when deriving the laws of Special Relativity one does not start with the energy mass equivalence, but with the observation that the velocity of light is independent of the reference frame. When doing so, the light plays from the beginning a special role.

When considering the complete theory one can note that photons are not fundamentally different from other particles. They are subject to the same laws as all the other particles. They are distinguished only by the values of those physical quantities which characterize them. As far as mechanics is concerned these quantities are the rest mass and the intrinsic angular momentum.

Another reason for the preferential treatment of the light may be that when introducing the Theory of Relativity kinematics is at the focus. Light flashes and light clocks in and at the side of running trains play an important role. In this way again the idea is conveyed that light is a special thing in relativity. This point of view can be understood when considering the situation at the beginning of the 20th century. At that time, nothing was known about gravitational waves that move with the same velocity as light. No neutrinos were known that move with almost the limiting velocity and there were no accelerators and colliders where many other particles are accelerated to the limiting velocity.

Disposal:

Say that there is a speed limit that is binding on all bodies and particles. Photons and gravitons move with exactly this velocity, as far as we know. It was believed for some time that this is also true for neutrinos.

Friedrich Herrmann

6.4 Velocity addition

Subject:

"When in […] u is eliminated, we get

$$u = \frac{u' + v}{1 + \frac{u'v}{c^2}}.$$

This is Einstein's relativistic law of velocity addition."

Deficiencies:

The equation tells us how the physical quantity *velocity* transforms when the reference frame is changed. There are corresponding laws for the transformation of length and time intervals, of energy and momentum, of electric and magnetic field strength, and others, but only in the case of the velocity one refers to "addition", instead of transformation. This may make the students believe that the case of velocity is basically different from the other transformations.

One should also remember that the term "addition" is reserved for the well-known mathematical operation.

When calling the above equation "velocity-addition formula" the students may believe that it is principally incorrect to add velocities in the normal way. Since normal mathematical addition can result in a velocity greater than c, the addition would not be allowed. However, this argument is not correct. One may and one must add up velocities, whatever the result of the operation is, when for instance, one wants to calculate an average velocity from many single velocities.

Origin:

Einstein himself called the formula addition-law [1]. The name is suggestive, since for small velocities the equation reduces to a simple mathematical addition.

Disposal:

Call the above equation transformation law instead of addition law.

[1] *A. Einstein*: Grundzüge der Relativitätstheorie, Akademieverlag Berlin, 1970, S. 39.

Friedrich Herrmann

6.5 The Michelson-Morley experiment

Subject:

"The introduction of a 'luminiferous ether' will prove to be superfluous as the view here to be developed will not require an 'absolutely stationary space' provided with special properties." (Einstein [1])

"There appears to be no acceptable experimental basis then for the idea of an ether, that is, for a preferred frame of reference. This is true wether we choose to regard the ether as stationary or as dragged along." [2]

Deficiencies:

The experiments of Michelson and Morley have shown that the velocity of light does not depend on the reference frame in which it is measured. The outcome of their experiment had several consequences for physics. One of them was of tremendous importance. It showed that a new theory was needed that embraces and modifies classical mechanics and electrodynamics. The other one has only indirectly to do with the first one: It was concluded that there is no luminiferous ether. Both consequences are often cited together, almost as if the non-existence of the ether was simply one of the many new statements of the Theory of Relativity. Actually, it is sometimes mentioned only by the way, as in Einstein's publication from 1905 [1]. Or it is assumed that the non-existence of a special reference frame is equivalent to the statement of the non-existence of a luminiferous ether, see our second quote [2].

We shall show in a thought experiment, that the two statements are independent from one another, and that one cannot be deduced from the other.

A car is running at a high velocity on a conveyor belt, which for the beginning is at rest. The velocity of the car relative to the belt is almost equal to *c*. We now turn on the motor of the conveyer belt, so that the belt moves in the same direction as the car. Although the sum of the velocities of the car and the belt is greater than *c*, we observe that the actual velocity of the car relative to the earth does not exceed *c*. If we run the conveyor belt in the opposite direction, the velocity of the car remains almost *c*. Suppose now, this experiment was done in place of the Michelson-Morley experiment. What would the experimenters have concluded? They would have concluded that there is a limiting value for the velocity and that it is not allowed to add velocities when changing the reference frame. These conclusions would have leaded to a new theory, the Theory of Relativity. This theory explains the observed results, which initially seemed so strange. Notice, that in no case the experimenters would have concluded that the carrier of the car, i.e. the conveyer belt does not exist. However, it was precisely this conclusion that had been drawn from the outcome of the real Michelson-Morley experiment. From the fact that the velocity of the light does not change upon a change of the reference frame, it was concluded that the carrier of the light wave does not exist.

Origin:

As long as there was no Theory of Relativity, the conclusion that a luminiferous ether does not exist seemed to be the only way out of the dilemma. With Einstein's theory however, the problem got solved in a completely new and unexpected way. We may consider it a gaffe that Einstein himself declared that the ether was dispensable. Some years after his first publication about his new theory he saw that he had not been right: "…there is no empty space, i.e. a space without a [gravitational] field." [3] Somewhat later he pronounced it even more clearly: "According to the General Theory of Relativity space without an ether is unthinkable." [4]

Disposal:

Leave the ether aside as long as you do not really have to do with it. Otherwise you will soon get entangled in the jumble of the concepts ether, space, gravitational field and vacuum. If you cite Einstein in connection with the ether, do cite his later declarations.

[1] *A. Einstein*: Zur Elektrodynamik bewegter Körper. Annalen der Physik und Chemie, Jg. 17, 1905, S. 891-921

[2] *R. Resnick*: Introduction to Special Relativity. New York: John Wiley & Sons, Inc., 1968, p. 33

[3] *A. Einstein*: Über die spezielle und die allgemeine Relativitätstheorie. Berlin: Akademieverlag, WTB, 1973, p. 125

[4] *A. Einstein*: Äther und Relativitätstheorie. Berlin: Verlag von Julius Springer, 1920, p.12

Friedrich Herrmann

6.6 Dilatation, contraction, expansion

Subject:

The concepts length contraction and time dilatation

Deficiencies:

In common speech and also in the technical language the words contraction and dilatation denote processes, i.e. something that happens as time goes on. Something that initially is long becomes shorter, it contracts; something that is short at the beginning becomes longer, it dilates or expands. This is not meant, however, when referring to length contraction or time dilatation in relativistic physics. Here, the value of a length and a time interval change only because the references frame is changed. In other words: because one choses another mathematical description for the same object or process.

Such a change of the values of physical quantities upon a change of the reference system is in physics the rule.

Every physicist knows that the value of the kinetic energy changes upon a change of the reference frame. But one would not say that the energy has *increased*. Such a statement would induce the question of what is the rate of change *dE/dt* of the energy.

Also the value of the momentum changes upon a change of the reference system. But one would not say that the momentum has increased, since then one would have to answer the question of what is the rate of change $d\mathbf{p}/dt$ and which is the force that causes it.

One might argue that these are subtleties: Everybody knows what is meant. This would be true if we had not to do with the theory of relativity.

With difficulty the beginner learns length contraction and time dilatation and what it is all about. Nothing with the contracted yardstick has changed and nothing with the clock, that runs slower from the viewpoint of other clocks; since upon changing the reference frame not only the yardstick has shortened but the whole world, and that is why one does not notice any contraction when living in this world.

But then the beginner learns about other subjects of relativity where something becomes shorter or longer, and he or she will believe that this is a phenomenon of the same kind as when changing the reference frame. For instance the Michelson interferometer with which gravitational waves are detected. The distance between the mirrors changes and thus the length of the light path and the interference pattern. But do not also change all the other lengths: the wave length of the light, the yardsticks, the body size of the researchers? If this were true the change of the length should not be detectable.

Of course, this conclusion is not correct. We have to do with a real process, which does not disappear when describing the situation in another reference frame. This insight is hampered or hindered if the two concepts „physical process“ and „change of the reference system“ are not clearly distinguished.

In addition, the problem is aggravated by the repeated asseveration that an ether does not exist or that space is empty. If space were empty in the sense that a normal mind imagines emptiness, the idea of an expansion of space would be meaningless.

Origin:

The denomination length contraction was appropriate within the context of the ideas of Lorentz [1]. According to Lorentz's theory length contraction was a real reduction of the distances in material objects. Already in the conception of his predecessor Fitz Gerald [2], Fig. 1, the contraction was real process.

LETTERS TO THE EDITOR.

****Correspondents are requested to be as brief as possible. The writer's name is in all cases required as proof of good faith.*

The editor will be glad to publish any queries consonant with the character of the journal.

Twenty copies of the number containing his communication will be furnished free to any correspondent on request.

The Ether and the Earth's Atmosphere.

I HAVE read with much interest Messrs. Michelson and Morley's wonderfully delicate experiment attempting to decide the important question as to how far the ether is carried along by the earth. Their result seems opposed to other experiments showing that the ether in the air can be carried along only to an inappreciable extent. I would suggest that almost the only hypothesis that can reconcile this opposition is that the length of material bodies changes, according as they are moving through the ether or across it, by an amount depending on the square of the ratio of their velocity to that of light. We know that electric forces are affected by the motion of the electrified bodies relative to the ether, and it seems a not improbable supposition that the molecular forces are affected by the motion, and that the size of a body alters consequently. It would be very important if secular experiments on electrical attractions between permanently electrified bodies, such as in a very delicate quadrant electrometer, were instituted in some of the equatorial parts of the earth to observe whether there is any diurnal and annual variation of attraction, — diurnal due to the rotation of the earth being added and subtracted from its orbital velocity; and annual similarly for its orbital velocity and the motion of the solar system. GEO. FRAS. FITZ GERALD.

Dublin, May 2.

Fig. 1. Prerelativistic idea about length contraction

Disposal:

Be careful to avoid the impression that Lorentz's formula describes a process. In particular, avoid the denomination length contraction and time dilatation.

On the contrary, when treating the interferometer for the detection of gravitational waves pronounce clearly that the distance between the mirrors changes; proceed correspondingly when treating the expansion of the universe. Here, the terms contraction or dilation are appropriate. The question will come up who or what is contracting or expanding. If you have banned the word ether from your vocabulary you need another name for the expanding entity. One does not do a favor to the student when calling it space. In the normal language space means no more than space for something. If the idea of space is no more than that, space can only increase but not expand.

Friedrich Herrmann

[1] *H. A. Lorentz:* Die relative Bewegung der Erde und des Äthers (Zittingsverlag, Amsterdam, Akad. v. Wet., 1 (1892), p. 74

[2] *G. F. Fitz Gerald:* The Ether and the Earth's Atmosphere, Science, Vol. XIII, No. 328, Letters to the editor, p. 390

6.7 Special Relativity and change of reference frame

Subject:

Special relativity is known to be a difficult theory. It requires from us to consider space and time not as two independent entities, but both together as one single concept. When going from one reference frame to another temporal as well as spacial intervals are changing.

Deficiencies:

There is no doubt that the Special Theory of Relativity (STR) is difficult. The reason is not mathematics, since it doesn't require from the learner more than the square root; lower secondary school mathematics is sufficient. The difficulty of STR must have another cause, and this cause seems to be obvious: the merge of space and time.

My experience from many discussions with beginners in the field as well as with colleagues who have treated the subject in their courses or lessons makes me suspect something else.

Let us begin with a little detour. Imagine it is the last Saturday of October, six o'clock in the evening. In the coming night daylight saving times will terminate. Somebody says: "Tomorrow at the same time it will be dark already." Somebody else says: "Tomorrow at the same time it is only five o'clock." Who of them is right?

Hardly anyone is able to judge the correctness of the two statements without thinking for a while. But why? None of the difficulties that mathematics or physics provide can be made responsible: no vector analysis, no differential equations, no curved space, no uncertainty relation. The reason why we stumble is the change of the reference frame. In the STR we have to do with the same kind of difficulty but to a greater extent.

In order to treat mathematically a problem of physics, in particular of mechanics or electrodynamics, a reference frame has to be chosen. A possible cause of confusion arises always when the reference frame is changed in the process.

The famous twin paradox is an example. In principle the physical treatment of the situation is simple. However, countless articles have been written to analyze it. In a nicely written book about relativity [1] that intends to present relativity at the level of the lower secondary school, the treatment of the twin paradox takes 11 pages.

The difficulty arises for the following reason: One first solves the problem in a reasonably chosen reference frame, i.e. that of the twin who does not travel to the remote star – which is very easy. However, one then insists to describe the solution in the reference frame of the second twin, which is a really unintelligent choice, since this reference frame is not free floating, but corresponds to a gravitational field that changes in time.

Also in classical physics nobody (perhaps apart from Ptolemaeus) would chose such a reference frame.

Origin:

The SRT initiated from the requirement that the laws of physics and the velocity of light should be independent of the reference frame. However, SRT was the origin of results that go far beyond the question of what happens when the reference frame is changed. The fact that we still begin relativity with the claim that the velocity of the light is invariant upon a change of the reference frame shows once more the congealing of the teaching contents.

Unfortunately, the name of the theory points to the change of the reference frame. It had been noted rather early that this name was not a good choice.

Disposal:

The most important equations of the (special-) relativistic dynamics can be derived in a few lines from the requirement of the identity of mass and energy. No change of the reference frame is required.

Let us remind on this occasion a rule that every physicist respects when describing a problem mathematically:

- choose the reference frame at the beginning in such a way that the description is simplest;
- don't change the reference frame in the course of the calculation.

Instead of titling the corresponding chapters "Theory of Relativity" call it better "Spacetime physics" as Taylor and Wheeler do [2].

Friedrich Herrmann

[1] G. Beyvers und E. Krusch: *Kleines 1x1 der Relativitätstheorie*. Books on Demand GmbH, Norderstedt, 2007, S. 67-77.

[2] E. F. Taylor and J. A. Wheeler: *Spacetime Physics*. W. H. Freeman & Co Ltd. (1992)

6.8 Mass, rest mass, invariant mass, relativistic mass, energy, rest energy and internal energy

Subject:

The Minkowski norm of the four-momentum that is invariant upon a change of the reference frame is called “invariant mass” or simply “mass”, in particular among particle physicists. Elsewhere it is also called “rest mass” or, when given in energy units, “rest energy”. Occasionally it is also called “internal energy”. Otherwise, in physics one understands by mass not only the invariant part, but the total mass, that changes its value with the velocity, which is also called “relativistic mass”. When measured in energy units it is simply called “energy”.

Deficiencies:

There are physical quantities, or better: names of physical quantities, that make trouble: They change their meaning over the course of time, or they are used by different persons with a different meaning. For some of these quantities the problem exists since a long time, or has always existed – for instance for the names “force” and for “heat”. Regarding mass the problem is more recent. For a long time, i.e. about 200 years it belonged to the more benign quantities. The chaos was provoked by the Theory of Relativity. The issue is simple, but the chaos is great.

Here the facts: There are two well-known equations

$$E^2 = E_0^2 + c^2p^2$$

and

$$E = mc^2$$

and the question is which names to use for the three quantities m, E and E_0.

In fact, with the discovery of the identity of mass and energy one of the two names mass and energy had becomes superfluous. But actually, several new names had been created, with the result that was described above.

Origin:

The problem originated with the come up of the Theory of Relativity. On the one hand there was the discovery that mass and energy are the same physical quantity: energy has the same properties as mass, namely gravity and inertia.

On the other hand a feature of the new theory is that it describes the physical world with four-vectors and its Lorentz-invariant norms. Lorentz invariants are convenient. They contain the essential of a particle or a process, they contain what is independent of the arbitrary choice of a reference frame. Since mass has stood for centuries for something that is characteristic for a particle, that represents an essential part of its identity, that does not depend on the reference frame, one let this name play this role also in the future. Thus, the name is used (in particular by particle physicists) for the Lorentz-invariant norm of the momentum four-vector, i.e. for what initially was called rest mass.

One can say that two requests are competing:

- the name mass as a measure for the inertia of a body or a particle (which can be great in one reference frame and small in another);
- the name mass for a quantity that characterizes a particle and that is independent of the reference frame.

Thus, the chaos is pre-programmed.

Those who use the word as a measure for inertia, needed a new name for the value of the mass in the center-of-mass frame. The designation “rest mass” and “rest energy” were obvious candidates. However the word “rest” must not be taken too seriously. It only means that the center-of-mass is at rest. Aside from this the system can display any amount of unrest.

Those who use the word “mass” for the Lorentz invariant, had to invent a new name for the measure of inertia. It was called “relativistic mass”. Since it might be feared that somebody does not know that by mass they mean the quantity m_0, they sometimes take the precaution to call it the “invariant mass”.

Disposal:

We do not dare to make a suggestion to the community of the particle physicists. However, regarding school we make a recommendation. Introduce mass as a quantity that characterizes the inertia and the gravity of a body. This concept is easy to understand. Later when relativity is introduced the students learn that mass and energy are one and the same quantity; they learn gravity and inertia increase when a body is heated, when a spring is tended, when a capacitor is charged. It is a matter of course that we use for this quantity the same name as before, namely “mass”. So we have not to revise our idea about mass. Once again: mass measures gravity and inertia.

One may call the quantities E_0 and m_0 “rest energy” and “rest mass”, even though nothing is at rest apart from the center of mass. The name “internal energy” might be more convenient.

Friedrich Herrmann

6.9 GPS correction and GTR

Subject:

The relativistic deviations of the clocks in a GPS satellite that must be corrected are of two kinds: One of them stems from the “time dilatation” due to the velocity of the satellite. The other one is due to the fact that the satellite is at a higher gravitational potential than the terrestrial clocks. The first effect is a special relativistic effect (STR effect in the following), the second, so it is often claimed, is explained only by the General Theory of Relativity (GTR). Sometimes it is said that the satellite clock’s advance is due to a weaker gravitational field.

Deficiencies:

Two problems arise with the effect that is ascribed to the GTR.

1. It has nothing to do with the field strength, but only depends on the potential. It is present also in the approximation of a homogeneous field, i.e. a field whose field strength is independent of the height.

2. The claim that the effect is an GTR effect is blundering. Certainly, one can argue about what kind of effect belongs to GTR. Is the fact that one takes seriously the equivalence of gravitational mass and inertial mass a GTR statement? Or do we use GTR as soon as we change into an accelerated reference frame? Rather not. It is more convenient to define the border between GTR and non-GTR as follows: Every phenomenon or effect that can be described with a flat Minkowski space is not an GTR effect.

If we adopt this criterion the above-mentioned difference in the proper times of two clocks at different heights is not an GTR effect. It can be observed also in a homogeneous field and its Riemann tensor is zero. In other words: the field can be transformed away by describing the situation in a free falling reference frame.

Consider the famous example of the twins (A and B), one of which (A) lives at the top of a high-rise building, the other (B) at the bottom. They meet half way up to adjust their clocks. After living for a while at the top and bottom respectively they meet again and compare their clocks and find that the clock of A indicates more than that of B. It is easy to explain this difference of the proper times by considering the building together with the twins in a free floating reference frame. Suppose that at the space-time point of the first clock adjustment a third person C jumps up in such a way that she is back at the space-time point of the second encounter of A and B. Whereas C is free floating or falling, the twins A and B depart and come back in an accelerated movement. The difference of the proper times between the two events for A and B can now be determined with the means of STR. The effect is the same as that of the classical twin paradox where the twin travel on different world lines from one space-time point to another.

Origin:

Who claims that the effect depends on different field strengths may argue: The fact that one clock displays more than the other must have a local cause; thus, there must be something at the two locations which is different for A and for B. But this argument shows that the concept of space-time has not been understood.

The argument in favor of an GTR effect might be: Whenever the gravitational field comes into play, STR is no longer valid. The field can only be transformed away by going into an accelerated reference frame. Accelerations, so may be the belief, do not belong to STR.

Disposal:

Treat both effects within the framework of the STR.

Friedrich Herrmann

6.10 Movement through spacetime

Subject:

„[The Geodesic Hypothesis] is the hypothesis that small 'freely-falling' bodies move along geodesic trajectories…"

„...as Krikalev hurtled along at 17 000 miles an hour onboard the Mir space station, time did not flow at the same rate for him as it did on Earth."

„When mass –be it a star, a planet or a human being– is present, spacetime bends around it so that an object traveling nearby must follow a rounded trajectory that takes it closer to the mass. Just as it is impossible to move in a straight line on the surface of a sphere, it is likewise impossible to move in a straight line through curved spacetime…"

Deficiencies:

The fact that one describes the world no longer in three-dimensional space, where time is only a parameter, that allows to order or align the various states of a system, but in spacetime, where space and time make up a whole, should have consequences for our way of speaking. In the colloquial language, which is always the basis for the description of physical phenomena, the separation of space and time, as described by classical physics, is firmly anchored. We speak about incidents that *happen*, objects that *move* and events, that are *the cause of later events*.

All this no longer works when taking spacetime seriously. When using the customary language to describe processes in spacetime we must be prepared to cause confusion.

Our citations show it in several ways. We discuss them one after the other.

1. In classical physics as well as in the world that we perceive with our senses a moving body has a trajectory. The trajectory is a curve in space. The forth dimension, i.e. time is taken into account by saying that the body *moves*. The body cannot move in spacetime. If one says that it moves on a world-line one is with one's mind already back on the three-dimensional trajectory. In spacetime the concept of movement has no sense. The same is true for the concept of current, that we normally imagine as the collective movement of a substance or in the case of the current of a physical quantity, as a movement of an imagined substance.

2. In physics, a rate of change refers to a time interval. So the rate of change of the electric charge is $\Delta Q/\Delta t$. Our second citation mentions a rate of change of the time itself. But what is this rate in the space station and what is it on the Earth. All we can do in this context is to divide a proper time interval by the coordinate time interval. But would that justify the claim that for Krikalev time flows at a higher rate for him than it did on Earth?

In everyday speech it is common to say that time flies by or that time passes quickly or slowly. However, this is not physics but psychology.

3. In our third citation we have again the movement through spacetime, see item one.

In addition, the author claims that it is impossible to move on a straight line through curved spacetime. But what would a straight line be? The geodesic is straight! If we follow two lines that are at first parallel and near to each other and we observe that their mutual distance increases or decreases, this can have two causes: Either the lines are curved or space is curved. In the case of the surface of a sphere it would be convenient to say that the great circles are straight lines in a curved space. (Of course it is possible that the lines and the space are both curved, and it can happen that their effects cancel each other. An example are the circles of latitude of the Earth: the distance between two of them is the same everywhere, although the lines are curved.)

Origin:

We have to do with a theory that destroys the basic categories of our description of the world. We cannot reproach the linguistic conflicts to the scientists who have developed the new theory. It is not their job to bring their theory in a shape that is appropriate for teaching to beginners. In our opinion teachers and publicists are not really aware of this problem.

Disposal:

Our everyday language fails, whereas the language of mathematics works well.

It seems that the only way out is to express oneself cautiously. That means:

1. Do not say: „Body K moves on a the worldline AB", but „AB is the worldline of body K".

2. Between two clock comparisons for Krikalev more time has passed than for his colleague an the Earth. Or alternatively: The clock that measures the coordinate time indicates more than that which measures the proper time. Or: The pointers of the proper-time clock move faster than the pointers of a coordinate time at the same position

3. Call a geodesic (in the two- or three-dimensional space) a straight line, because in its space it is not curved.

Friedrich Herrmann

6.11 The relativity of simultaneity

Subject:

When the special theory of relativity is treated in a schoolbook, the relativity of simultaneity is an important subject, besides length contraction and time dilatation. Often a take home message is formulated: “Simultaneity is relative.”

Deficiencies:

We believe that too much importance is given to the subject in comparison with other propositions of the special theory of relativity (STR). It is rather intricate but there is hardly any consequence for something that is relevant.

The question of whether two events, that take place at two different places are simultaneous results from our conviction or experience that there exists a time that is independent from position and velocity: a parameter that allows us to order the states of the world as a whole unambiguously. In order to be able to answer the question, one would need a procedure that allows to decide about if two events at two different positions are simultaneous or not. This is done by defining a procedure to “synchronize” clocks at the various positions.

To take a step back, let us ask another, but similar question: Is equal-positionness also relative? Or in more fluid language: Do two events, that for one person happen at the same place or position, also take place at the same position for another person? With „for one person“ and „for another person“ we mean „in one reference frame“ and „in any other reference frame“. The answer to this question is „no“. This is so obvious that nobody would even ask the question.

That also the statement about the simultaneity is not really significant is best seen when looking at it from the view point of the general theory of relativity (GTR). There, the question vanishes like sand between the fingers. The concept of simultaneity looses its sense, since it is in general not possible to synchronize two identical clocks, unless one decides give a new meaning to the word synchronization. For example the clocks of the GPS system: In the satellite one installs a clock that would run slower if placed next to a normal clock at the surface of the Earth. When installed on the satellite, this clock runs „synchronous“ to the terrestrial clocks: Each time the two clocks meet, they indicate the same time. But attention: This type of synchronous running is not meant, when one talks about synchronization in the context of the STR.

Like length contraction and time dilatation, the relativity of simultaneity has to do with the change of the reference frame. Changes of the reference system are often the cause of confusion, not only in relativistic but also in classical physics, and not only in the minds of students but also in those of experienced physicists [1]. When changing the reference frame we always have to keep in mind: It is not the nature, not the real world that changes. What changes is only our mathematical description. Only slightly exaggerating one might say: The change of the values of the physical quantities upon a change of the reference frame, and the resulting changes in the interpretation of a phenomenon are the result of an inadequacy of our description. But unfortunately we have no alternative.

But why does the relativity of simultaneity have no important consequences? Because the relation between two events that are simultaneous in one reference frame and not simultaneous in another one are not causally connected. Thus, the inversion of the temporal order cannot have any consequence.

Origin:

In his famous paper “On the electrodynamics of moving bodies” [2] Einstein treats the subject synchronization right at the beginning at great length over three pages. It is the first problem the reader who is interested in Einstein’s ideas is confronted with.

At the time he wrote the article Einstein could not yet suspect the strange fate that the concept would suffer by a theory that he himself developed in the following years.

Whereas in 1905 it seemed to be a justified concern, to discuss a very human expectation, namely that one is able to decide if two events that take place at two different locations are simultaneous or not, independent of the reference frame, from the viewpoint of the GTR this effort appears a desperate attempt so save a misconception of our intuitive view of the world.

Disposal:

The disposal brings us a gain of time: Simply refrain from treating the subject, at least at school. There are other, more important results and statements in the context of spacetime, that usually come off badly.

Friedrich Herrmann

[1] *F. Herrmann:* Altlasten der Physik, Relativitätstheorie und Bezugssystemwechsel

[2] *A. Einstein:* Zur Elektrodynamik bewegter Körper, Annalen der Physik und Chemie, Jg. 17, 1905, S. 891–921.

6.12 The name “Theory of Relativity”

Subject:

The two great theories of Einstein are called tTheories of Relativity: the Special and the General Theory of Relativity (STR and GTR). They are based on the principle of relativity: the laws of nature have the same structure for all observers. In the STR, the principle applies only to reference frames that move uniformly against each other. In the GTR, it is generalized to accelerated reference frames.

Deficiencies:

The term theory of relativity suggests that changes of the reference frame play a special role in Einstein’s theories. It also gives the impression that changing the reference frame is the main subject of the theories. This is reflected in the teaching, especially in the case of the STR. Before coming to the interesting statements of the theory, one has to work through the intricate considerations of mutually moving frames of reference, with the result that the learners (in school and college) quickly lose their interest in the subject.

Think about what you would tell about the STR to someone who you know is not ready to listen for more than two minutes. I do not believe it would be reasonable to tell him or her that the principle of relativity applies. Here are some better suggestions:

- Space and time merge into one entity.
- Instead with three-vectors nature is described with four-vectors.
- Energy and mass are the same physical quantity.
- There is an upper limit for the speed.

It may be objected that the name of a theory has no influence on the student’s understanding of it. I disagree. In my experience, teaching success depends strongly on the language, and especially the terms that are used. Nomen est omen: If the teacher, on the basis of the name, has come to the conclusion that the main purpose of the theory of relativity is the description of changes of the reference frame, then he or she will structure the lecture accordingly.

Those who know the theories well may not understand these concerns. But it’s not the professionals I’m worried about. Rather, it is those for whom „relativistic“ physics is no more than one of many other educational topics. What is kept in mind is the somehow tricky behavior of lengths and time intervals in the case of a reference frame change. (The equation $E = mc^2$ is known anyway, because it is encountered on graffiti, book titles or stamps.) The essence of the theory actually comes under the wheels.

Origin:

The name was coined very early. In 1906, Planck first switched from the term “Lorentz-Einstein Theory” to “Relative theory”. It soon became “Theory of relativity”, which was also used by Einstein in 1907.

Imagine that the course of the story was a little different: that the STR originated in a different way, for example, from the experimental observation that energy has inertia and weight, so that energy and mass are found to be the same physical quantity. What would you call the theory? Maybe equivalence theory? Certainly a very different teaching tradition would have established.

Disposal:

Instead theory of relativity or relativistic physics, give the corresponding chapter a different title, such as Wheeler for instance: Physics of spacetime.

Or, if one can decide not to start the subject with kinematics, but rather with dynamics, the title might be: *The velocity limit*, or *The identity of mass and energy*.

Friedrich Herrmann

6.13 Teaching the twin paradox?

Subject:

The twin paradox is discussed in high school textbooks, university textbooks and popular science books. Hundreds of articles have been published in scientific journals. Even "meta-articles" have been written, i.e. articles that try to classify the scientific papers that have appeared so far and to present their appearance in a histogram. It is only a small topic, but it is obviously considered important.

Here a short reminder: Two twins – in our case called Willy and Lilly – are together and synchronize their watches. Then Lilly makes a long journey with a spacecraft at constant speed to a distant star; she then turns back and travels – again at constant speed – back. When Lilly and Willy rendezvous again, they discover that more time has passed according to Willy's watch than to that of Lilly. Willy has aged more than Lilly. Figure 1 shows the path-time diagrams of Willy and Lilly, the so-called world lines.

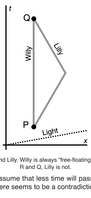


Fig. 1. World lines of Willy and Lilly. Willy is always "free-floating" between space-time points R and Q, Lilly is not.

If both Willy and Lilly assume that less time will pass for the other because of the "time dilation", there seems to be a contradiction.

Deficiencies:

I assume, dear reader, that the paradox is familiar to you. I am not interested in resolving it. God knows, this has been done often enough.

Rather, I am concerned with the role that the subject plays and should play in the teaching of school and university.

To this end, I would first like to tell the story in a slightly different way.

Willy and Lilly compare their watches. They display the same readings. Then Lilly goes to the *playground*, and Willy goes shopping. In the evening they meet again and find that their watches no longer match, Fig. 2.

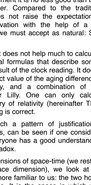


Fig. 2. Lilly goes to the playground, Willy goes shopping. Between the two space-time points, time passes differently for them.

Of course, the story is unrealistic: the clocks do not run accurately enough. But as a thought experiment it is no less good than the usual story, which is not very realistic either. Compared to the traditional one, it has the advantage that it does not raise the expectation that one can better understand the observation with the help of a calculation. Rather, it expresses a fact that we must accept as natural: Space and time form a unity.

For an understanding it does not help much to calculate the time difference with the help of several formulas that describe something that is no less implausible than the result of the clock reading. It does not help much if one tries to justify the correct value of the aging difference of the two twins by a time dilation for Willy, and a combination of time dilation and an acceleration effect for Lilly. One can only calculate that within the framework of the theory of relativity (hereinafter TR), which one must of course know, everything is correct.

How inappropriate such a pattern of justification and explanation is, especially for beginners, can be seen if one considers a rather analogous situation for which everyone has a good understanding, but which nobody would present as a paradox.

Instead of the two dimensions of space-time (we restrict ourselves in TR, as usual, to a single space dimension), we look at two other dimensions whose relationship is more familiar to us: the two horizontal components of normal positional space. It is the space in which we constantly navigate around as beings bound to the earth's surface.

Here's our story: Willy and Lilly drive, each with a car, on different paths from a point P to a point Q, fig. 3. Willy drives straight ahead, directly from P to Q. Lilly almost always drives straight ahead, except that her path has a kink. Both read their respective speedometers at the beginning and end of the journey. They notice that Lilly has covered a longer distance.

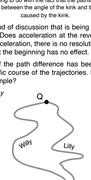


Fig. 3. Willy and Lilly drive on different paths from P to Q. Their odometers show that the covered distances are not the same.

If we now assume that we were not aware that the two dimensions forward and sideways are two dimensions of one and the same space, then the following paradox would arise (in analogy to the TR twin paradox) Willy notes that Lilly has to travel a greater distance to get as far forward as he does, Fig. 4a.

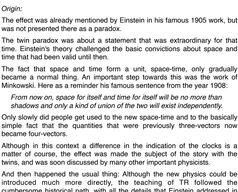


Fig. 4. (a) Willy realizes that Lilly has to drive a longer distance to get ahead. (b) Lilly realizes that Willy has to drive a longer distance to get ahead.

Apparently, her path is subject to a length dilatation. But Lilly comes to the same conclusion, fig. 4b. From her point of view, Willy has to travel a greater distance in order to advance as far as she has. So: for each of them the distance of the other is longer. That would be the paradox. Of course, not both can be true. And if they look at their speedometers, they also realize: For Willy the conclusion was right: Lilly is covering a longer distance. Lilly's conclusion was wrong.

Now, if one were to discuss the problem in the same way as the real twin paradox, one would examine the question of what role the turn of Lilly's path at point R plays, and what changes when the turn is not a *sharp* kink, but a slightly *gentler* arc, and so on. One would find that although the kink in Lilly's trajectory is necessary to interpret the observation, a kink does not necessarily cause a large difference in the path length in general. In Fig. 5, Willy's trajectory has three kinks, but Lilly's is still the longer one.

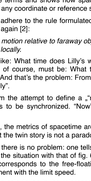


Fig. 5. The kink has something to do with the fact that the paths are of different lengths. But there is no simple relation between the angle of the kink and the lengthening of the path caused by the kink.

This is precisely the kind of discussion that is being held in connection with the real twin paradox: Does acceleration at the reversal point play a role? Yes and no. Without acceleration, there is no resolution of the twin paradox, but acceleration right at the beginning has no effect.

Back to Fig. 3: Even if the path difference has been calculated, it is only known for a very specific course of the trajectories. But what about the two paths in Fig. 6, for example?

Fig. 6. One wants to make someone understand that the paths of Willy and Lilly do not have the same length. Would one calculate the path lengths using differential geometry?

Would one calculate the distance Willy has travelled by adding up the progress perceived by Lilly, but then taking into account what effect Willy's change of direction has on the result at any given moment? Probably not. But then how do you get the answer? Simply by measuring it locally, i.e. with the speedometer of the respective car. And of course, one would not be surprised that the displayed values are different.

In the same way we will describe the situation of fig. 2: We measure the "increase" of time locally, with a watch on board.

Origin:

The effect was already mentioned by Einstein in his famous 1905 work, but was not presented there as a paradox.

The twin paradox was about a statement that was extraordinary for that time. Einstein's theory challenged the basic convictions about space and time that had been valid until then.

The fact that space and time form a unit, space-time, only gradually became a normal thing. An important step towards this was the work of Minkowski. Here as a reminder his famous sentence from the year 1908:

> *From now on, space for itself and time for itself will be no more than shadows and only a kind of union of the two will exist independently.*

Only slowly did people get used to the new space-time and to the basically simple fact that the quantities that were previously three-vectors now became four-vectors.

Although in this context a difference in the indication of the clocks is a matter of course, the effect was made the subject of the story with the twins, and was soon discussed by many other important physicists.

And then happened the usual thing: Although the new physics could be introduced much more directly, the teaching of TR followed the cumbersome historical path, with all the details that Einstein addressed in his first work on the subject: clock synchronization, relativity of simultaneity, length contraction, time dilation.

Every student must go this way, and inevitably encounters the twin paradox.

And finally, something else may also play a role: the belief that one can only gain an understanding by calculating.

Disposal:

The whole problem is solved if it is made clear from the outset that space and time form a unit, space-time. Wheeler, for example, explains the reality of space-time in simple terms and shows how space-time can be defined unambiguously without any coordinate or reference system [1].

First of all, one should adhere to the rule formulated by Wheeler, which he reminds us of time and again [2]:

> *Don't try to describe motion relative to faraway objects. Physics is simple only when analyzed locally.*

Thus avoid questions like: What time does Lilly's watch indicate for Willy. Because the question, of course, must be: What time does Lilly's watch indicate for Willy *now*? And that's the problem: From the "now" for Willy you can't infer a now for "Lilly".

One should refrain from the attempt to define a „now" for distant places with the help of clocks to be synchronized. "Now" should only be used "here".

In an advanced course, the metrics of spacetime are introduced. Also here it becomes obvious that the twin story is not a paradox.

But even for beginners there is no problem: one tells the story that belongs to fig. 2 and compares the situation with that of fig. 6, taking as experience that the longest time corresponds to the free-floating movement, and no time at all to the movement with the limit speed.

In no case one does the calculation in the reference system of Lilly. This is because it violates a rule that is respected everywhere else in physics: Choose your coordinate or reference system in a way that makes it as easy as possible to handle the problem, and most importantly: Don't change it in the middle of the calculation. But that's exactly what one does when formulating the twin paradox, and that's what causes the chaotic discussions that go along with it.

[1] J. A. Wheeler, *A Journey into Gravity and Spacetime*, The Scientific American Library HPHLP, New York, 1990, Chapter 3

[2] C. W. Misner, K. S. Thorne, J. A. Wheeler, *Gravitation*, W. H. Freeman and Company, New York, 1973, p. 4.

Friedrich Herrmann

6.14 Longitudinal and transverse mass

Subject:

In the treatment of special relativity, sometimes a longitudinal and a transverse mass is introduced. This is to express that the inertia of a body is different (greater) in the direction of movement than in the direction transverse to it.

Deficiencies:

The need to introduce two new mass concepts arises, if one insists, that mass should be a measure for the inertia. In fact, the inertia of a body moving at relativistic speed is greater in the direction of motion than transversely.

Two remarks in this regard:

1. Irrespective of whether the mass does us the favour of measuring inertia or not, we want to ask ourselves the question what to understand by inertia in the context of a process of movement. It is reasonable to define an inertia T as follows:

$$T := F/a \qquad (1)$$

and this always, i.e. not only in the case of classical movements where the force is proportional to the acceleration, i.e. when

$$T = m\ .$$

We bring equation (1) into another form. With $a = dv/dt$ and $F = dp/dt$ we get

$$T := dp/dv$$

The inertia defined in this way tells us how much momentum dp must be supplied to a body so that its velocity changes by dv.

Since we know the relativistic relation between p and v, we can easily calculate the inertia. For a change of momentum in forward direction we find

$$T_l(v) = \frac{m_0}{\left(1-\frac{v^2}{c^2}\right)^{\frac{3}{2}}}$$

and for the transverse direction

$$T_t(v) = \frac{m_0}{\sqrt{1-\frac{v^2}{c^2}}}$$

Let us first have a look at the inertia in forward direction. It is neither identical with the rest mass nor with the relativistic mass

$$m(v) = \frac{m_0}{\sqrt{1-\frac{v^2}{c^2}}}$$

This is easy to see when looking at the $p(v)$ relation, Fig. 1. T is given by the slope of the curve, i.e. the differential quotient dp/dv, see the red tangent to the curve. The relativistic mass, however, is equal to the slope of the green straight line. Only at the beginning, in “classical approximation”, the slope dp/dv is equal to p/v, and thus equal to the rest mass, see the blue tangent.

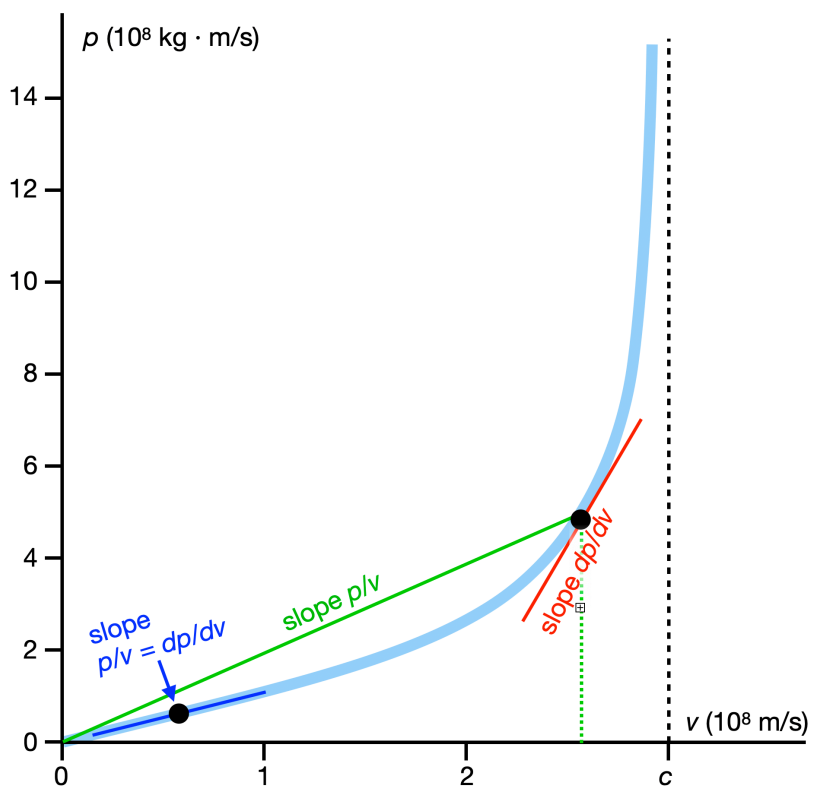


Fig. 1. The inertia of a body is given by the slope of the function $p(v)$. It depends on the velocity.

Now to the transverse inertia: It is that of a body which does not move in the transverse direction. However, this does not mean that it is described by the rest mass, since the mass of the body has increased due to the high longitudinal velocity.

In short: inertia is a quantity, which has a greater value in a given, well-defined direction than in the orthogonal direction or in other words: it is a tensor.

2. Should we conclude that there exists another tensorial mass besides the rest mass and the relativistic mass? This is not a good question. A physical quantity exists, if we introduce it, if we define it. Let us try to ask the question in a better way: Should we introduce a tensorial inertial mass in addition to the rest mass and the relativistic mass? A cautious answer would be: We should do so, if it is useful, if it is worthwhile. And is it worthwhile? The answer to this question is probably rather: No.

But isn’t it a pity about the beautiful interpretation of mass as a universal measure of inertia?

A pity perhaps – but why should mass be better off than other physical quantities? Let us remember:

- When we construct or invent a new theory, we are happy if the variables it contains measure simple properties known to us from our everyday experience. Most of the time, however, this does not quite work. Think of force, for example, or heat.
- The inertia behaves similar to some electrical quantities. The resistance characterizes an object: a resistor. If somebody says that the resistor has a resistance of 10 kΩ, then one is informed. However, this is only possible if the current is proportional to the voltage. But what if it is not? How do we characterize for example a semiconductor diode? In this case it is not enough to give one number. One has to give the U-I characteristic curve. The same applies to the capacitance. And we are in the same situation with the inertia. Inertia cannot be described by a single number; one needs a characteristic curve, Fig. 1.

Origin:

The concepts of longitudinal and transverse mass were introduced by Lorentz in 1899 and they were also calculated by Einstein in 1905 using his theory of relativity. Since then, they have been haunting physics, although they have no apparent use.

Disposal:

With the rest mass and the relativistic mass there are enough masses, not to mention the possibility to introduce consequently a longitudinal and a transverse energy. Nothing is missing if the longitudinal and transverse masses are ignored. The fact that a body has different inertia in the forward and transverse directions can be accommodated in an exercise, but introducing two new terms would be a bit too much of a good thing.

All that is to be understood in this context is contained in the diagram of Fig. 1. It becomes even clearer, if one does not, as usual, plot momentum versus velocity, but rather velocity versus momentum, Fig. 2, because as independent variable one chooses, if possible, that quantity, on whose values one has the most direct influence – and that is not velocity, but momentum. We push the accelerator pedal so that the engine pumps momentum from the earth into the car, and see on the speedometer what consequence this has, i.e. what velocity results from it.

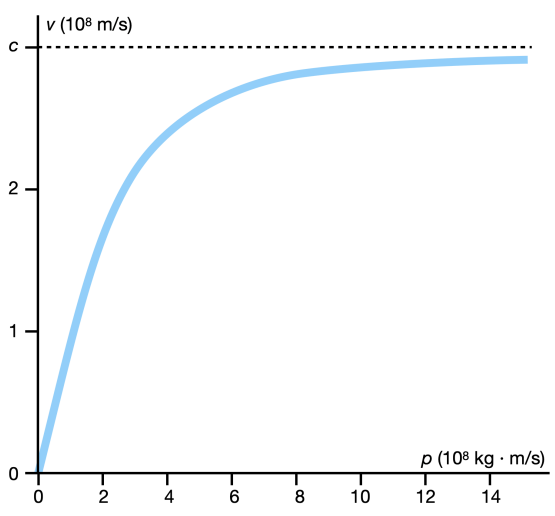


Fig. 2. The function $v(p)$ tells us everything about the inertial behavior of a body.

But what happens then to the nice rule that mass is a measure of inertia? Well, we have to relativize that a bit: It measures inertia only as long as the speed is not too high. Only for $v << c$, inertia is an intrinsic property of a body, and does not depend on its state.

Friedrich Herrmann

6.15 Absolute spacetime

Subject:

The absolute space does not exist, so we are assured again and again. Speaking about the position or the movement of a body only makes sense in relation to other bodies.

Deficiencies:

As is known, the absolute space was introduced by Newton. It manifests itself in the fact that one can decide whether a body rotates without having to refer to another body. The body rotates in the space or against the space. A problem was that there was no way to decide whether a body was doing a translational motion with respect to space. It would have been necessary that there be some kind of "milestones" in the space. The only milestones that were known were the other bodies, and that is why one later insisted that positions and velocities are relative; that it only makes sense to speak of position and velocity relative to other bodies. Ernst Mach expresses it particularly clearly: the absolute position and the absolute movement, i.e. the movement against the absolute space are only "mental constructs" [1].

The only way out of the dilemma seemed to be, from the point of view of that time, a gigantic action at a distance: A body moves relative to the celestial sphere (at Newton's time) or relative to the fixed stars (at Mach's time). For Newton, such an idea was unacceptable.

But it got worse, and Mach could not have known that yet either: In his time, the universe consisted only of our own Milky Way; nothing was known yet about the other galaxies and galaxy clusters, and nothing was known about the expansion of the universe. Since one has this knowledge, the idea of the action of the fixed stars on a body her, where we are, has become even more absurd. Against whom should a body that is now here with us move? Against the position of the stars now, or against their position then, when they emitted the light that we receive now. And what is meant by "now" for a distant star?

Finally, the year 1915 brought the solution: Space, which for Newton, and also for Mach, still appeared completely homogeneous and structureless, is not homogeneous and structureless at all. It has local properties, and these properties are different from one location to another. They are expressed by the metric tensor (or by the Riemann tensor), whose components are functions of position and time. So Newton's absolute space is back again, except that it does not have the property assumed by Newton of being inalterable, and thus of not containing any milestones.

Origin:

It was already addressed. One should not have taken the shortcoming of the Newtonian space that tragically. There was no contradiction, but only a lack. The milestones were already present at that time, only one could not see them yet. Newton's ideas, however, were the right ones. The positivistic attitude of Mach is a useful basic attitude for the scientific work, but often it hinders the imagination. It is also worth remembering Mach's rejection of the atomic theory, which originated from the same basic attitude.

Thus, after Goethe, once again someone had failed in his attempt to disprove Newton.

Disposal:

As early as when treating Newtonian mechanics, one introduces the space as something really existing, against which the movements take place. One addresses properties, at first without deepening the idea of space curvature. What is important is only that the idea of space as a concrete, real existing entity is created. The question whether the space is absolute or not, does not arise then any more.

Of course, later it will be explained that space and time together make up the entity called space-time.

One might be inclined to declare: So there is an absolute space-time. But this is also superfluous once it has been clarified that the spacetime has local properties, whereby local means: local in the four-dimensional spacetime.

However, this does not solve one minor problem: the name. Colloquially, the term "space" means: Space for something. It therefore does not refer to something that exists. Space means: something could be there. As an alternative, one might think of reactivating the nice old designation ether. However: apart from the inglorious past of the concept, the name ether has the disadvantage that it clearly refers to the content of space. The ether is like a gas, for example. But the gas also needs the space in which it is.

The space, with which we were here, the space of the general theory of relativity, is container and content at the same time. There is no container without space and there is no space without container. And for this, we do not know any example, no analogue, no model from our empirical world.

This idea should be taught in the classroom. But what could we call it? So far, we have not come up with anything suitable. So let's stick with the term space, or space-time. And the students have to learn: In physics, space is not an empty container.

[1] *Mach, E*: Die Mechanik in ihrer Entwicklung. Leipzig: Brockhaus, 1897, S. 223

Friedrich Herrmann

7

Oscillations and Waves

7.1 Resonance frequency and natural frequency

Subject:

When treating driven harmonic oscillations, one usually emphasizes that the resonance frequency is not exactly but only approximately equal to the natural frequency of the oscillator.

Deficiencies:

What can we do with such a statement? Apparently nature was not able to arrange oscillations reasonably. First we learn that there is resonance when the oscillator is in time with the driving device. But then we are told that the resonance frequency and the natural frequency of the oscillator do not match exactly. Nature seems to spoil the game. Do we have to conclude that the original idea is not correct? An uneasiness comes up.

The incongruity can easily be dismantled. Resonance means that the energy which the oscillator absorbs and dissipates as a function of the excitation frequency has a maximum value. Since

$P = \mathbf{v} \cdot \mathbf{F}_0$,

this maximum is located on the frequency axis at the same position as the maximum of the velocity. (We assume that the oscillator is driven by a force with a constant amplitude $\mathbf{F}_0$. Similar arguments hold when the driving is realized with a constant velocity amplitude.) Now, the frequency that belongs to the maximum of the velocity amplitude is indeed the natural frequency. As a consequence, the position amplitude cannot not have its maximum at the natural frequency. Neither does the frequency of the acceleration amplitude coincide with the natural frequency.

From

$$x(t) = x_0(\omega) \cdot \sin(\omega t)$$

follows

$$\dot{x}(t) = \omega \cdot x_0(\omega) \cdot \cos(\omega t) = v_0(\omega) \cdot \cos(\omega t)$$

and

$$\ddot{x}(t) = -\omega^2 \cdot x_0(\omega) \cdot \sin(\omega t) = a_0(\omega) \cdot \sin(\omega t)$$

If the velocity amplitude $v_0(\omega) = \omega \cdot x_0(\omega)$ had its maximum at the frequency ω_{res} , then neither the position amplitude $x_0(\omega)$, nor the acceleration amplitude $a_0(\omega) = -\omega^2 \cdot x_0(\omega)$ will have its maximum at this frequency. Thus, the discrepancy between the natural frequency and the “resonance frequency” is due to an inappropriate choice of the quantity that is considered. Obviously numerous other quantities could be displayed as a function of frequency, and the maximum will be found at various different positions on the frequency axis. From this observation one will not conclude that resonance takes place at different frequencies according to which quantity’s maximum is considered.

Origin:

Presumably our tendency to put in the foreground what we see with our eyes. We have become accustomed to regard a mechanical problem as solved when we know the trajectory of the bodies, i.e. the position as a function of time. But again and again we have to admit that in mechanics the quantities momentum and energy are more fundamental than the kinematic quantities.

Disposal:

Not define resonance by means of the positional amplitude, i.e. by the manifest quantity. Resonance is when the absorbed energy has its maximum value.

Friedrich Herrmann

7.2 Forced oscillations and phase difference

Subject:

In the context of forced mechanical oscillations students learn that at resonance there is a phase difference of $\pi/2$ between the driving mechanism and the oscillator. This result is often formulated as a key sentence, for instance:

"In the case of resonance, there is a phase difference of $\Delta\varphi = \pi/2$."

or:

"In the case of resonance the pendulum falls short of the exciting oscillation by a quarter of a period."

Deficiencies:

1. A phase difference always refers to two physical quantities with a sinusoidal time dependence. In the case of the driven harmonic oscillator it is often not said, which are the quantities to whom the statement refers. However, since only the position coordinates are considered anyway, nobody will ask for phase differences between other quantities. But one could have studied just as well other phase differences. So, one of the quantities could also have been the velocity, the acceleration, the momentum of the oscillating body, or the force that is acting on it. The second quantity could have been the position, the velocity or the acceleration of the driving mechanism. One could chose any two of these quantities and ask for the corresponding phase difference. Most of these phase differences are not easy to interpret, however, and that is also true for the phase difference in the cited propositions. What do we learn by knowing that the phase difference between the position coordinates of the driving mechanism and the oscillator is $\pi/2$?

2. A driven mechanical spring oscillator consists of the following components: a moving body, a spring and a driving mechanism. The fact that the oscillation is damped can be taken into account by adding yet a forth component, a damper (in the electric analogue this would be a resistor). These four elements can be combined in several different ways. The mechanical "circuitry" can have various different topologies (in the same way as the corresponding electric circuit could have). In order to define the behavior of the oscillator unambiguously we also have to dispose of the properties of the energy source, i.e. the driving mechanism. It is not enough to demand that the driver is sinusoidal. It has to be decided which (if any) amplitude remains constant when changing the frequency: that of the position, the velocity, the force or energy flow. The shape of the resonance curve depends on this choice. Among all these possible combinations there are two for which the problem gets particularly transparent:

– All of the four elements are connected in parallel and the force amplitude of the driving mechanism is held constant, Fig. 1;

– All of the four elements are connected in series and the velocity amplitude of the driving mechanism is held constant, Fig. 2.

(Also in the electric case these two basic circuits exist. When the electric elements are connected in parallel, the current amplitude has to be fixed, when connected in series, the voltage of the energy source is held constant.)

Now, the statements cited above are not valid for either of these basic circuits, but for a hybrid of the parallel and the series circuit. Correspondingly, the interpretation of the statement about the phase difference is somewhat difficult. On the contrary, in the case of either of the basic circuits the interpretation is simple. We shall discuss the example of the parallel circuit, Fig. 1.

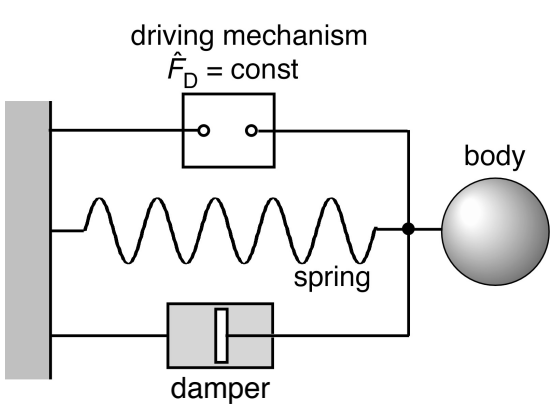


Fig. 1. Parallel oscillator has constant driving force amplitude.

Resonance means that the time average of the energy flow from the driving mechanism (index D) to the oscillator

$$\bar{P} = \overline{v_D F_D}$$

has a maximum value. With

$$v_D = \hat{v}_D \sin(\omega t)$$

and

$$F_D = \hat{F}_D \sin(\omega t - \phi)$$

we get

$$\bar{P} = \frac{\hat{v}_D \hat{F}_D}{2} \cos\phi\,.$$

In this expression all of the three factors, namely $\hat{v}_D$, $\hat{F}_D$ and $\cos\phi$ can in principle depend on frequency. For the "parallel oscillator", Fig. 1, the force amplitude $\hat{F}_D$ is held constant, it is independent of the frequency. Each of the other two factors have a maximum value at the resonance frequency. Thus for the resonance we have $\cos\phi = 1$, or $\phi = 0$. That means that the velocity of the driving mechanism and the force which it exerts on the oscillator are "in phase". This statement is plausible. In order to excite or drive an oscillator most effectively one has to push or pull most strongly when the oscillator moves most quickly.

For the series oscillator, Fig. 2, the velocity amplitude is frequency-independent. The force amplitude and $\cos\phi$ both have a maximum value at the resonance frequency and again we have $\phi = 0$.

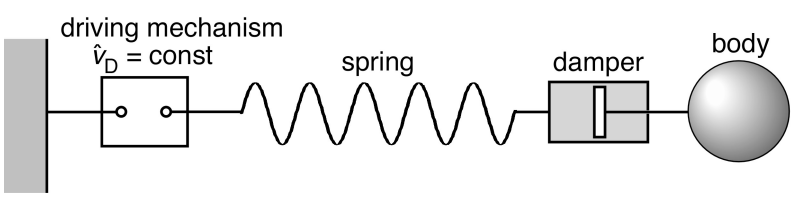


Fig. 2. Series oscillator has constant velocity amplitude.

The oscillator topology that most often is considered in mechanics is that of Fig. 3. It can be shown that this oscillator is mathematically equivalent to the parallel oscillator. For the force one has to write $D\,\hat{x}_D \sin(\omega t)$. In the case of resonance this force is in phase with the velocity of the oscillator. This fact can be justified in the same way as the zero phase difference for the parallel circuit. Since the phase difference between position and velocity of the oscillator is $\pi/2$, the statement of the citations at the beginning follows.

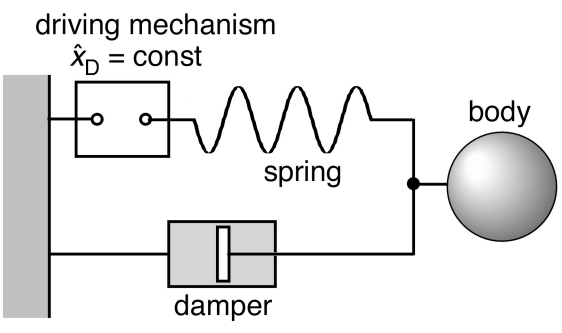


Fig. 3. Hybrid form of parallel and series oscillator.

Origin:

See our previous article [1]. Putting the position amplitudes of the driving mechanism and the oscillator in the center of attention corresponds to the tradition of mechanics to consider a problem as solved when the position-time relation is determined, i.e. when we have calculated what we see with our eyes. However, we understand mechanics better when putting in the fore the quantities momentum and energy and the respective flows.

Disposal:

Considering a phase difference as a function of the excitation frequency is worthwhile only if the corresponding function is interpreted. It is easy to interpret the phase difference between the force and the velocity. The product of them is the flow of the dissipated energy. A phase difference of zero contributes to make this product a maximum for the resonance frequency.

[1] *F. Herrmann*: Resonant frequency and natural frequency, article 7.1

Friedrich Herrmann and Holger Hauptmann

7.3 Huygens' Principle

Subject:

In contemporary physics text books Huygens' principle is not only used to explain the diffraction of light by a single slit, a double slit and a diffraction grating, but also the reflection and the refraction of a plane wave.

Deficiencies:

1. Huygens' principle (or the Huygens-Fresnel principle) is a simple mathematical tool for determining the interference pattern of two or more single waves. However, a particular principle is not needed in the case of the simplest and at the same time most important interference experiments. Even without Huygens' principle one will expect that a circular or spherical wave will emerge from a small opening (small compared with the wavelength) in an obstacle on which a plane wave is incident. There is no need for a new principle in the case that there are two or more such openings either. Moreover, there is no reason for a particular name "elementary waves" for the emerging circular waves. The principle is useful only when the slit is greater than the wavelength.

2. Also for the description of reflection and refraction Huygens' principle is not needed, since it explains the behavior of a plane wave by that of circular waves. A function can be decomposed in many different ways: in harmonic components, in spherical harmonics, Bessel functions and many more. If such a decomposition is done, it is reasonable to chose a basic set of functions that takes into account the symmetry of the problem. Obviously this is not the case when decomposing a plane wave into "elementary waves", i.e. circular waves. The original wave, i.e. the plane wave has already the highest symmetry that a wave can have. Reflection and refraction are easily understood with plane waves. Using spherical or circular waves means to explain the simple by the complicated.

Origin:

The principle was formulated by Huygens in 1690 in his "Traité de la Lumière". This was 100 years before the great age of wave optics which began with Fresnel and Young, and 150 years before the Electrodynamics of Faraday and Maxwell. In Huygens' time the laws of reflection and difraction were known, it was known that the velocity of light is finite, as well as the fact that light is composed of colored components. Why then was the principle at that time so important, and why did it keep its significance until today?

At Huygens' age another theory of the light existed already: The corpuscular theory, first advanced by Descartes and later by Newton. To this theory Huygens opposed his idea of light as a wave. The criterion for a good theory at that time was mainly its ability to explain refraction and reflection.

To explain meant (and still means today), to reduce a phenomenon to another one, that is taken for fundamental and thus not in need of explanation. However, since the time of Fresnel refraction and reflection do not need elementary waves as an explanation. When finally Maxwell showed that light is an electromagnetic wave and described it mathematically Huygens' elementary waves definitely became obsolete, even though it was not clear why it should be valid for the complicated electromagnetic transverse waves. Only Kirchhoff succeeded in showing the compatibility of Huygens' principle with the electromagnetic theory.

The role which Huygens' principle plays today at the school and the University is still marked by its former importance. Just as Lenz' law or Kepler's laws, it has survived its own more general follow-up laws. It is true that it still is a useful method for an approximate determination of interference patterns, but as such we should put it together with the many other tools of physics and not call it a principle.

Disposal:

To explain the diffraction at the single slit and the interference pattern of the double slit and the grating, no particular principle is needed. If one cannot decide to let the treatment of the large slit to the university, then one may introduce Huygens' principle, but with a more modest demeanor.

Friedrich Herrmann

7.4 Double slit diffraction and interference of light

Subject:

The single-slit and the double-slit experiment play an important role in the teaching of physics. The diffraction patterns are discussed extensively. The double-slit experiment is presented as a proof of the wave character of the light. Later it is used as a means to demonstrate the nature of so-called quantum objects.

Deficiencies:

1. The diffraction of light at the single slit and the double slit is treated in a detailedness and thoroughness, that exceeds the standards of a general education. When *Young* carried the double-slit experiment out for the first time it played the role of an *experimentum crucis*. Today we know such a great number of other proofs of the wave character of the light, that the double-slit experiment has lost much of its original significance. Moreover, we know that the light corresponds to only a small fraction of the electromagnetic spectrum. Regarding the other types of electromagnetic radiation, we care much less for a proof of their wave character. We simply take it for granted that they are waves – with good cause, since nothing would work if it was not so.

2. The single- and double-slit diffraction experiments are complicated. They combine two phenomena, which are not always clearly distinguished: diffraction and interference. Sometimes it is even claimed that there is no difference between them.

3. If we would ask an unbiased student to design an experiment which shows the interference of two light waves, he or she would definitely not think of the double-slit arrangement. The manifest idea would be to use two light sources. Only if the student has understood why such an experiment does not work, and even not when using two lasers, he or she will accept that a more sophisticated idea is needed. A well-designed experiment should give the student the feeling: “This could have been my own idea”. The double-slit experiment is surely not of this kind.

4. In quantum physics the double-slit experiment is used as the stage for all kinds of contradictory stories. Light is imagined as consisting of photons, i.e. tiny bodies, which, in order to get from the light source to the detector have to pass either through one or through the other slit. In spite of all warnings, that are pronounced, the idea of the individual tiny particles is ineradicable. To emphasize the particular character of the photons, they are now often called quantum objects, instead of particles. But even so, when speaking about them, the language remains that which is used when speaking about small individuals. Actually, as soon as the question of through which slit a photon is going, one has already admitted that one takes the idea of small bodies for legitimate. Simultaneously, one has made a statement about the size of the particles: Their lateral extension must be smaller than the slit width.

Origin:

1. Already before Young’s experiments and Fresnel’s theory, there were good arguments in favor of both, the wave and the particle model of the light. Naturally one was convinced that only on of them could be “true”: Either the light is a wave or it consists of particles. With Young’s experiments the verdict seemed to be rendered. In the following decades many more arguments in favor of light as a wave were found, but not only that. With Maxwell’s theory, 70 years after Young’s experiments, the nature of the waves was understood, or at least it was believed so. Moreover, more and more waves of this nature were fond on both sides of the spectral region of the visible light. In spite of all these new evidence for the wave nature of the light, when teaching we still attribute to Young’s experiments a significance as if it were the only proof of the wave character of the light.

2. Quantum physics led the double-slit experiment to new heights. We learn from quantum mechanics that we rejoiced too quickly. The “true nature of the light” is more intricate.

3. A stable tradition of problems in written exams has developed in which no final secondary-school examination and university-entrance examination can be imagined without a problem about the diffraction of a wave at a single or double slit.

Disposal:

1. When the inference of light wave is to be shown experimentally, first discuss thoroughly why the experiment cannot be done with two laser beams.

2. Instead of discussing the double-slit and the single-slit arrangement, use from the beginning a grating. The results are more convincing.

3. For some purposes the Michelson-Interferometer is more appropriate. The advantage: no sine function, no diffraction.

4. Discuss the diffraction phenomenon with radiations where the effect is much greater than with light: electromagnetic microwaves, or sound waves. Then the interesting question is not why light shows diffraction but on the contrary, why diffraction of the visible light is such a small effect.

Friedrich Herrmann

7.5 Coherence of waves

Subject:

In textbooks the concept of coherence is explained in various different ways. The following citations are taken from different books:

(1) “Wave trains which interfere with one another are called coherent, those which do not interfere are incoherent.”

(2) “Two wave generators, which produce a permanent interference pattern are called coherent. In order to do so they must oscillate with the same frequency and a constant phase difference.“

(3) “For an extended light source, e.g. a glowing filament, the wave trains emanating from different points of the filament and striking the eye are incoherent, i.e. they have completely different phases and directions of polarization.

(4) “Only light which starts from one point of a light source, can be brought to interfere, after being splitted and traversing different ways.“

(5) “Since the light that is spontaneously emitted by a hot body is radiated from atoms that are independent from one another, it is excluded that two different light sources incidentally execute the same oscillation, i.e. emit coherent wave trains.“

(6) “A slit emits coherent light as long as for its width d and for its angle of aperture light cone 2α holds:

$d \cdot \sin \alpha < \lambda/2$.”

Deficiencies:

Not only high school students but also university students have problems with the concept of coherence. The definitions cited above show that this is no wonder. Some of them are hard to understand by themselves. But things get particularly difficult when trying to reconcile these statements with one another.

In the following, the numbers refer to the numbers of our citations.

What is the object to which a statement about coherence refers? According to the citations (1), (3) and (5) it refers to the relation between two “wave trains”. But what is a wave train? The whole wave? Or part of it? Which part?

According to definition (2) coherence expresses the relation between two wave generators. It is said, that these have to oscillate with the same frequency and a constant phase difference. Does that mean that there are oscillators that can oscillate with the same frequency and a phase which is not constant?

Citation (6) attributes the coherence simply to the light.

Now, the question is if these definitions are only different formulations of the same fact or do some of them contradict one another?

Definition (3) tells us that only light which emanates from one point is coherent. Definition (4) makes a similar statement. But what is meant by two different points? Is there a maximum distance which is allowed? Definition (5) says it more clearly: Light which comes from different atoms cannot be coherent. However, it is well-known, that light from a distant star is used to determine the star’s diameter by means of Michelson’s stellar interferometer. In this case light interferes which comes from sources that can be a million km distant one from the other.

Origin:

All the sentences (1) to (6) make statements either about how to create coherent light or how to demonstrate coherence. Non of them tells us what is the nature of coherent light. But if we know only the property or nature of the source, how can we judge the coherence of a light field whose sources are unknown or unspecified, for instance the water waves on the ocean?

Here, we note the tendency to describe the generation process or the detection process of a phenomenon instead of the phenomenon itself. Usually these processes are more complicated than the real phenomenon. To understand how a bicycle works, we do not need to know the production process in the bicycle factory. In order to understand what a sound wave is, we do not need to know the working principle of an organ pipe or the human hearing.

Another cause for some incongruities is the tendency to consider a phenomenon as understood only when it is reduced to a statement about the behavior of some particles. Coherence is a phenomenon which can perfectly be described by means of classical wave theory. When looking for an interpretation in the context of quantum phenomena one easily gets trapped in the brushwood of models and interpretations.

Disposal:

Let us begin with two general remarks concerning the concept of coherence:

1. Coherence, which can be more or less pronounced, is a property of the light. It is understood that the light owes its properties to a light source. But that does not mean that coherence or incoherence is a property of the source.

2. Coherence is a local property of the light. That means that a given light distribution can be more coherent at one place than at another. So the spacial coherence of the light that is emitted by a star is minimum at the star’s surface and is almost perfect (maximum) here at the Earth i.e. at a great distance from the star.

When we say that coherence is a local property, we do not mean that coherence can be attributed to a point in the sense of mathematics. (In this sense no physical quantity is local.)

Coherence can be explained or defined in various ways. It manifests itself in each theory which is used to describe the light: geometrical optics, classical wave optics, the thermodynamics of light and quantum electrodynamics. Since it is our goal to explain the concept to a beginner, we will choose the simplest of these theories, i.e. geometrical optics. After this we will hint at how this explanation translates into wave optics. We advise not to try an explanation on the atomic scale at the school. This is a subject for the University.

We limit ourselves to evaluate the degree of coherence qualitatively. Let us try to describe the light in a small domain of space just in front of us. Which kind of light rays are crossing this space? We consider four situations which are particularly simple.

We are in the middle of dense fog. Our space domain is crossed by light rays of all directions. The light is a mixture of light of all spectral colors, indicated in Fig. 1 by differently dashed lines.

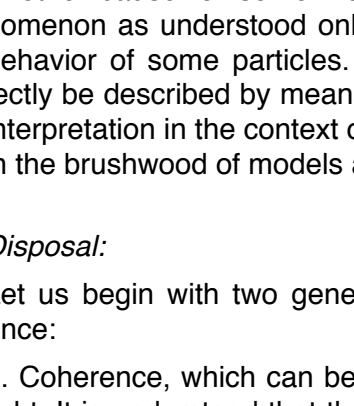


Fig. 1. All colors, all directions. The light is temporally and spatially incoherent.

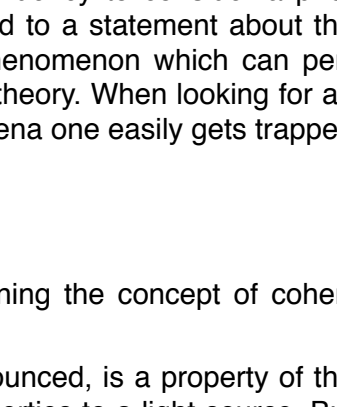


Fig. 2. One single color, all directions. The light is temporally coherent.

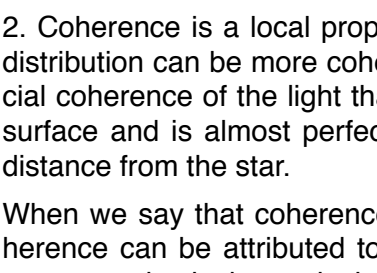


Fig. 3. A single direction, all colors. The light is spatially coherent.

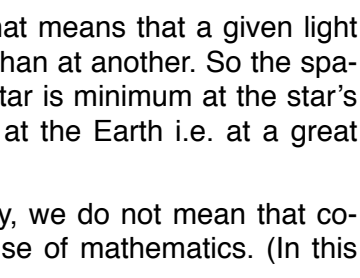


Fig. 4. A single color, a single direction. The light is temporally and spatially coherent.

Next we assume it is night, dense fog again and there is a street lamp that emits monochromatic light. Again the light in our space domain comes from all directions, Fig. 2.

In our third situation it is night, no fog, no moonlight, no starlight. At a great horizontal distance there is a incandescent light source. Now all light rays which cross our space have the same direction, but it is light with many different spectral colors, Fig. 3.

Finally a situation similar to the one before, but with a lamp that emits monochromatic light, Fig. 4. Now all rays have the same direction and all the light has the same spectral color.

The light in fig. 1 is completely incoherent. That of fig. 2 is called “temporally coherent”. Thus temporal coherent light is monochromatic light. The light of Fig. 3 is spatially coherent. Thus, spatially coherent is the opposite of diffuse. The light of Fig. 4 finally is temporally and spatially coherent.

Here yet an analogy or allegory that one may tell to the students. We consider a crate with apples. The apples have a great range of colors and sizes. We want to classify them. We begin by assorting them according to their size into 10 different boxes, each box for a given size interval. Now in each box the apples are uniform with respect to one of our criteria, i.e. size. Next we assort the apples of each box in one of 10 smaller boxes according to color. Altogether we now have 100 boxes. In each of these boxes the apples are uniform with respect to both our criteria size and color.

The similarity between apples and light goes even further. It is obvious that we can get uniform apples from the initial crate only by sorting out those apples that do not correspond to the desired size and color. It is not possible to transform the multi-color multi-size apples into uni-color uni-size apples, in the same way as it is not possible to transform incoherent light into coherent light – which would mean destruction of entropy and thus violate the second law.

One can, on the contrary, grow trees that produce uniform apples from the beginning. In the same way we can employ a light source that produces coherent light from the beginning, i.e. a laser.

At the end a word about the wave optical description of coherence. Light is temporally coherent when the dispersion of the magnitude of the **k** vector is small, it is spatially coherent when the angular dispersion of **k** is small. It is easy to tell the coherence by looking at a waves, say for instance the waves on the surface of a lake. There may be sections of the wave field that look like sine waves. These sections have a certain lengths and a certain widths. The length is a measure of the temporal coherence, the width is a measure of the spatial coherence.

Friedrich Herrmann

7.6 Electromagnetic transverse waves

Subject:

Right at the beginning of the chapter about waves students learn the definition of the concepts longitudinal and transverse wave:

“For a transverse wave the displacement of the individual sections of the wave carrier is perpendicular to the direction of propagation. For a longitudinal wave they oscillate back and forth in the direction of propagation.”

Later, when the subject is electrodynamics, they learn:

“Light can be polarized. Thus, it is a transverse wave whose ***E*** and ***B*** fields oscillate perpendicularly to the direction of propagation.“

Usually, the distribution of the electric and the magnetic field strength in space is illustrated by a figure like that of our Fig. 1.

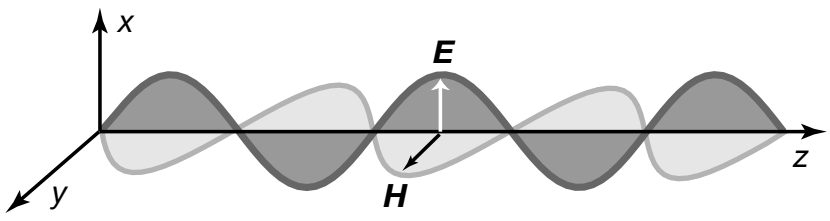


Fig. 1. “Snapshot” of the electric and the magnetic field strength of a sine wave

Deficiencies:

According to the definition which our students learn, in a transverse wave the wave carrier *moves* perpendicularly to the direction of propagation. If we take this definition literally, then an electromagnetic wave is not a transverse wave, since nothing is moving in such a wave. Of course, one might argue that the statement is not meant literally, but just in the way we speak normally when we say that the temperature or the stock-market price “is moving”.

However, the “movement” seems to be taken too seriously by the students. We suspect that part of the fault is the picture of Fig. 1 which is never missing in the text books: A snapshot of the movement of the vector tip of the electric and the magnetic field strength.

You can easily find out that something is not understood when performing a physics examination at the University. Ask for the field line picture within the room where the examination takes place for the radio waves coming from a nearby radio station, the students usually reply by sketching the picture of Fig. 1. When you point out that this is not a field line picture, the students are usually perplexed. Apparently, they interpret the image of Fig. 1 in the sense of our citation: a movement. What makes the interpretation of the figure somewhat difficult is the fact that first a spacial coordinate system is drawn, and then two other physical quantities ***E*** and ***B*** are represented. We know the procedure from mechanics, where we often draw force vectors in a scene that represents an object in normal space. In our case, there is the additional difficulty that the values of ***E*** and ***B*** change from point to point, and that their functional dependency is shown for only one space coordinate. The suggestion of an oscillation in the sense of a movement is rather strong.

Origin:

A somewhat unreflected take-over of the definition of the concepts longitudinal and transverse wave from mechanics to electrodynamics. There may exist a historical reason why the oscillation metaphor is so widely used in electrodynamics. In former times students learned: “Light is a transverse ether wave.” And that was meant in the sense of the mechanical definition of the concept transverse wave.

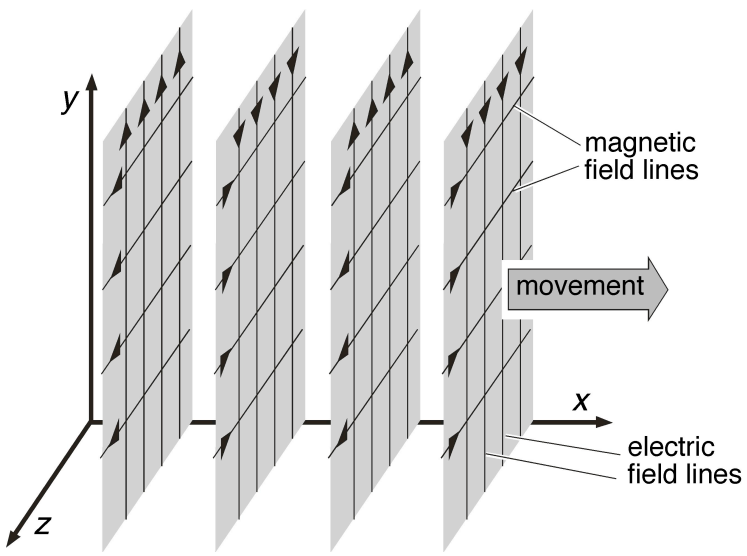


Fig. 2. Field line picture of a periodic electromagnetic wave

Disposal:

Explain the field strength distribution in a (periodic) wave with a drawing like that in Fig. 2, instead that of figure 1.

Friedrich Herrmann

7.7 Unpolarized light

Subject:

What is meant by the denotation “unpolarized light”? The following citations try to give an answer.

“The ***E*** field vectors of the light wave oscillate in no preferential direction. One refers to polarization when the ***E*** field vectors move in a well-determined manner. White light is in general unpolarized.”

“In general electromagnetic radiation is the superposition of a great number of single waves with a different orientations of the oscillation planes and with different phases.”

“Natural light is in general unpolarized. It originates in atomic transitions of a great number of atoms. Each atom emits a light wave, whose direction of polarization is statistically distributed in space, so that the plane of oscillation of the emitted light changes steadily.”

Sometimes, unpolarized light is represented by a picture like that of figure 1, which apparently is supposed to be a snapshot of the electric field strength (more exactly: of the tip of the vector arrow) above the position coordinate in the direction of the propagation of the light. One can see various “waves” at the same time at the same place.

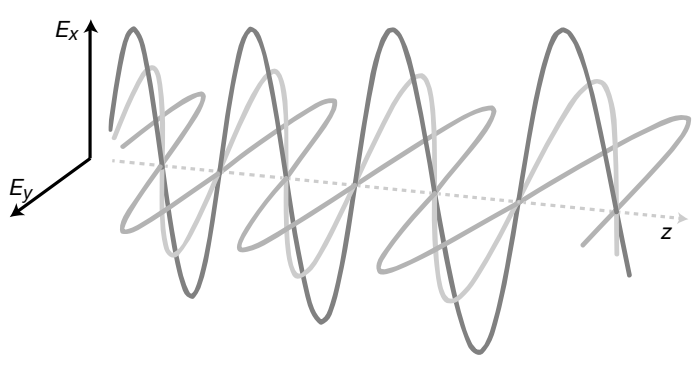


Fig. 1. “Snapshot” of the electric field strength vector tip in an electromagnetic wave. Is the wave unpolarized?

Deficiencies:

It is not difficult to understand the concept of a polarized electromagnetic wave. Neither is it difficult to understand how a polarizer works. The question of how we can imagine unpolarized light seems to be more difficult. In school books this question is somewhat neglected.

There are several theories of the light: geometrical optics, classical electrodynamics, quantum electrodynamics and thermodynamics. Depending on which of these theories is applied, the explanation of what is unpolarized light is somewhat different. Here we shall limit ourselves to classical electrodynamics.

The state of polarization of a light beam is best described by making a statement about the behavior of the electric field strength vector in a plane perpendicular to the propagation direction as a function of time; or in graphical representation: about the movement of the vector tip in this plane. (We admit that the light beam is homogeneous in its transverse extension.)

Light can exist in various states of polarization. The most important and best-known are linear polarization, elliptical polarization (with the special case of circular polarization) and the complete absence of polarization.

In the case of linearly polarized light the vector tip describes a harmonic movement, in the case of elliptic polarization an elliptic movement. There are many other possibilities for preparing light in such a way that the vector tip executes a more or less regular movement, among them Lissajous curves. When the light is unpolarized the vector tip moves on irregular curves without any periodicity. The average velocity of this movement depends on the temperature of the light and the length of the vector arrow depends on the intensity of the light. Both the direction and the module change irregularly. We could also describe the vector by its cartesian components. Then we would say: Both the *x* and the *y* component of the vector vary irregularly. In both descriptions there are two contributions to the “disorder” of the state of the light and thus to the entropy that is transported by the light beam.

Regarding our citations:

1. The first citation says, that the ***E*** field vectors oscillate, and that they do not have a preferred direction. Usually by oscillation we understand a periodic process. However, if the light is white, the vector tip does not make a periodic, but an irregular movement.

2. The second citation says that thermal radiation is a superposition of single waves. This statement goes a little far. First one should specify what is meant by “single wave”. One might believe that a sine wave is meant. Then the single waves would simply be the harmonic components of the light. If that is meant, it would be more appropriate to say that the radiation can be decomposed into such components, just as it can be decomposed in many other ways. But it may be that the harmonic components are not meant. Our third citation gives an indication.

3. “Each atom emits a light wave, whose…”. Here we see, that the light wave cannot be a pure sine wave. Since it originates in one atom it has a beginning and an end. According to a conception that many students have, such a “light wave” is an object that can be individually identified or at least imagined. Here probably the photon is haunting around, but in a somewhat vulgarized form: a small object which resembles a piece of wire that had been given a wavy shape. It conserves its individuality even when it is part of a light beam. Some pictures in text books foster such an idea.

4. Text books often show pictures that illustrate the working principle of a polarizing filter. Sometimes these pictures are like that of Fig. 1. Here three “individual waves” are shown. They have the same wavelength and are in phase. The figure does not show how long they are. When considering only that part which is represented, the superposition results simply in a linearly polarized wave. The idea that the state of the wave is one with a maximum of disorder (of entropy), cannot be seen from the figure.

Origin:

It seems that the problems has several causes.

1. When the students learn that light is a transverse wave, they may believe that the tip of the field strength vector must oscillate in a direction perpendicular to the propagation of the wave.

2. A tendency to believe that a wave *consists* of spectral components instead of seeing these components as a result of our arbitrary decomposition. The wave seems to consist of them like a book consists of pages.

3. A somewhat naive idea about the photon. Light consists of these individuals, but radio waves do not.

4. The awe to consider the light under a thermodynamical point of view.

Disposal:

White light that is completely incoherent is omnipresent. So, do not hesitate to describe the field strength distribution of such light and discuss the various contributions to the disorder of this state.

Do not speculate about the “true nature” of the light. Remain committed to what we know: How to describe the state of polarization and of absence of polarization by means of electrodynamics. Some thermodynamics in the arguments is not harmful.

Avoid the word oscillation when describing unpolarized light. The field strength vector does not oscillate; it moves chaotically.

Friedrich Herrmann

7.8 Tuning fork and resonance box

Subject:

"When the foot of a struck tuning fork is brought in contact with a resonance body or a table top or even the cranial bone, the produced sound is amplified and much easier to hear."

"A tuning fork is struck and put in contact with several objects. Sometimes the sound becomes louder. it is loudest for a tunig fork with a resonant box."

"The sound waves produced by an oscillating tuning fork are very gentle. A hard underlay serves as resonance body for the tuning fork, so that the oscillation is amplified and becomes audible.

Deficiencies:

Already in the famous text book by Pohl [1] one can read: "Often it is said that 'the oscillations are amplified by resonance'. This is a rather weird way of speaking." Pohl's wrote this a long time ago, but apparently what he said did not get around in the meantime. Our quotations, that are typical, show it: Still today one frequently hears that the sound is *amplified* by a resonance body, or simply that it becomes louder. This is not really incorrect, but indeed weird, as Pohl expresses it. It sounds as if something could be got for free.

The statement is similar to the following: If we spend much money, the turnover increases. At first it sounds good. But the problem is seen here more easily: If your turnover in the first week of the month is high, it may be that nothing is left for the remaining three weeks.

The situation is similar with the sound: with the resonance box the sound of the tuning fork is louder, but it lasts correspondingly less time. Because of the emission of sound by the box the oscillation of the box is strongly damped, and the tuning fork is damped by the resonance box.

This is similar to the electric circuit of Fig. 1.

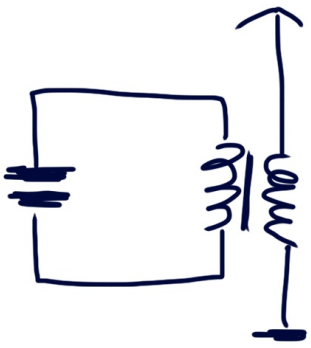


Fig. 1. The oscillating circuit at the left is first charged with energy. If it was not coupled to the antenna oscillator at the right, it would oscillate for a long time. Because of the coupling it looses its energy rapidly. The energy goes away with the emitted electromagnetic wave.

The oscillating circuit at the left taken alone is only weakly damped. Once charged with energy it would oscillate for a long time. However, it is coupled to the antenna oscillating circuit at the right. The antenna oscillator is strongly damped because it emits an electromagnetic wave. Because of the inductive coupling the first oscillator looses its energy quickly to the second, so that it oscillates only for a short time.

It is not appropriate to speak of an "amplification" in this context, since normally the word is used with a different meaning in science and technology: In an amplifier a signal enters with a small energy current, and it comes out with a great energy flow. In order to achieve this the amplifier has to be connected to an energy source.

The resonator box on the contrary only ensures that the energy of the tuning fork goes away quickly.

Origin:

The sensual perception is put in the fore-ground instead of the balance of the conserved quantity energy.

Disposal:

Explain that the resonance box ensures that the energy is quickly released with the emitted wave. The energy flow is greater, and the sound is louder but is lasts a shorter time than without the resonator box.

Friedrich Herrmann

[1] *R. W. Pohl:* Mechanik, Akustik, Wärmelehre, Springer-Verlag Berlin (1969), S. 235.

7.9 Coupled pendulums, coupled oscillations and synchronization

Subject:

Titles in physics textbooks:

- „Coupled pendulums“
- „Coupled oscillations“

What is meant is the same in both cases: Two pendulums or spring oscillators are coupled by means of an elastic spring.

Deficiencies:

The *pendulums* in Fig. 1 are coupled. To say that the *oscillations* are coupled is awkward. Indeed, insight is of the experiment is that we have to do with two independent oscillations, or two movements that are not coupled. In physics, one talks of a coupling whenever a system cannot be decomposed in two sub-systems that do not interact. In other words: when the Hamiltonian, or more generally the Gibbs-Massieu function does not consist of two summands with no common variables.

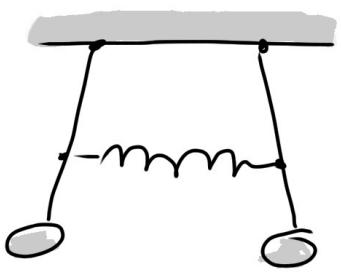


Fig. 1. Coupled pendulums, but not coupled oscillations

By choosing appropriate coordinates the system can be decomposed in two subsystems that do not interact. Each of the two coordinates describes one of the two normal modes.

The tendency to consider the two physical pendulums in the first place instead of the decoupled normal modes is also seen in the traditional explanation pattern of the so-called *injection locking*, a surprising phenomenon, discovered by Huygens: Several pendulum clocks of identical design are mounted in a common housing. After a while they oscillate synchronously and with a will-defined phase difference. At first, it appears as a miracle that one pendulums allows its neighbor to tell it with which frequency and phase it has to oscillate. Doesn’t each of them have its own favorite frequency? How can this be influenced by the other pendulum?

To answer the question the customary explanations resort to a somewhat beamy tool: The process is non-linear. Such an approach is comprehensive and correct, but also unnecessarily intimidating. It does not take into account a useful rule for the teaching of science: Explain a phenomenon by considering the simplest case where it shows up. Finally, when treating the normal familiar oscillations we also proceed in this way: First the undamped harmonic oscillation. If there is time left over we continue with the damped, the forced, the self-exited, the non-linear and the relaxation oscillation.

Origin:

Our tendency to base our physical description on what we see with our eyes. In the present case the movement of the individual pendulums.

Disposal:

We do not refer to coupled oscillations. If we absolutely want to use the term „coupled“ then we talk about coupled pendulums. But we could also say: a harmonic oscillator with two degrees of freedom.

Regarding the phenomenon of synchronization (injection locking): We consider the simplest example where the phenomenon shows up: a spring oscillator where two massive bodies are coupled by a spring, Fig. 2. If the system is excited in any way, the two bodies will in general make an irregular movement.

Fig. 2. Spring oscillator with two degrees of freedom

We then introduce a damping, represented in Fig. 3 by the dashpot symbols. Thereby, in general the two normal modes will be damped differently with the result that one of them will die away faster than the other. If we had to do with a self-oscillation, as in the case of Huygens’ clocks, the system will absorb energy preferentially in time with the less-damped mode with the result that this mode will be maintained whereas the other one will even lose energy.

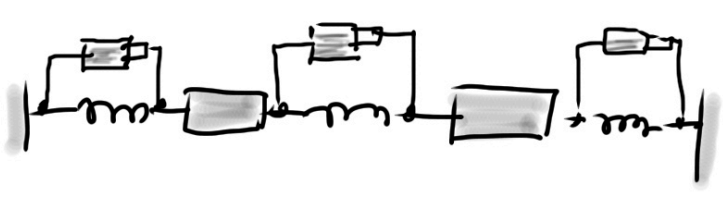


Fig. 3. Spring oscillator with damping. One mode is more strongly damped than the other one.

This behavior appears to be a synchronization: If we consider the two bodies as the partial systems they seem to make an agreement about a common frequency – a kind of miracle. If, on the contrary we focus on two normal modes it turns out that the process is no more than a dying off of one of them. Such a behavior is to be expected. It would be very improbable that both of them are equally damped.

Friedrich Herrmann

8

Atomic and Quantum Physics

8.1 The concept of trajectory in quantum mechanics

Subject:

“In quantum mechanics, due to Heisenberg's uncertainty principle, the notion of a particle trajectory does no longer make sense.” We all know the statement in these or similar words.

Deficiencies:

What does this statement mean? Does it mean that the concept of trajectory has a sense in every other context or in every other branch of physics which is not quantum mechanics? What, then, is the sense of a trajectory in thermodynamics? Which is the sense in geometrical optics? Or in wave optics? Which is the sense in our everyday life? Who or what can be expected to have a trajectory? What is the trajectory of a cloud? What is the trajectory of an amount of money that is transferred? Or of the data transmitted in the internet? That the concept of trajectory does not make sense is not an exception, it is the rule.

So why is it so remarkable that the concept does not make sense in quantum mechanics? Because we use an inconvenient model: the model of the point-like body that can be tracked individually. One then has one’s hands full with repairing the damage. Thus, the problem is home-made.

Consider a stationary state, for instance the ground state of the hydrogen atom. Neither the theory, nor the experiment tell us that that the electron is point-like. Neither the theory, nor the experiment tell us that something is moving. If we hadn’t talked our students into believing into the tiny objects whizzing around the disclaimer concerning the trajectory would not have been necessary.

Origin:

The particle model according to which all what happens in the physical world can be reduced to the movement of small, individual “particles” and their interaction was extremely successful until the turn of the 20th century.

It is understandable that one does not like to throw away such a powerful tool. However, in the following time the tool was also applied for purposes for which it was not suitable. It was adjusted or distorted until it lost its original force. In this way came up the probability interpretation of quantum mechanics, which for the common sense is so hard to accept.

Even though we physicists know it better, we steadily contribute to keep the inconvenient model alive. There is no examination where Bohr’s model is not asked. And even the student who does not know much, he or she knows Bohr’s model, of which we had shown that it is insufficient for our purposes.

There are many other instances where we tell or suggest to our students, what in principle we would like to white out. Although there are no books where it is written, students hear it very often: The electrons *move around the nucleus*. The students hear the incorrect statement more often than the correct one.

Disposal:

Trust in what quantum mechanics tells us. Its reliability is well established. Do not use models, which themselves are the cause of problems of comprehension.

Friedrich Herrmann

8.2 Illustrations of the atom

Subject:

Our subject is a picture, that is familiar not only to physicists. We will not reproduce it here for reasons which will become clear in a moment. It is the picture or illustration of the atom: the nucleus as a small spherical structure, surrounded by ellipses, the trajectories of the electrons, on which sometimes the electrons themselves are shown as small spherical bodies.

Deficiencies:

"A picture is worth a thousand words." It is easier to memorize than a verbal or a mathematical description of an object. Pictures are vital tools for the teacher. However, since they are so easily assimilated they can sometimes hinder an intended learning process. They do so when they describe a subject incorrectly or in a way that does not correspond to the intention of the teacher. They are pictures that are not chosen by the teacher, but which haunt the world and reproduce themselves. They may be so intrusive that nobody can escape them. Even those who know their harmful effect succumb to their suggestive power. An example is the image of the atom that was just mentioned and that corresponds to Bohr's model. We find it in books and journals of popular science, we run across it cast in bronze as a company logo on the main gate of the company, it is reproduced in millions of copies on stamps and on paper money, and it is also found surprisingly often in specialized physical journals.

What is to blame with these representations is that since quantum physics came into being they simply do no longer correspond to our idea of the atom. We spent a significant portion of teaching time in order to show the weakness of these images and to show why it must be replaced by another one. But in the minds of the students the picture is already engraved, so that we will have only a limited success. And what gets stuck in the minds after a longer period of time is only the circulating bodies of Bohr's model.

Origin:

The pictures came into being with the introduction of the model of the atom by Rutherford and Bohr. But less than 20 years later, when Schrödinger (and Heisenberg, Born and Jordan) invented quantum mechanics, it became obsolete.

Now, the first successful model has always an advantage over follow-up models. The younger model must displace the older one. But in physics such a process was rarely successful.

Disposal:

Bohr's model of the atom and the corresponding pictures are a very interesting subject for the history of physics. In the teaching of physics, however, they are counterproductive. We can try to compete against the bodies flying on elliptic trajectories only with counter-images that are more beautiful and evocative, like for instance colored density plots of the psi-square distribution.

Friedrich Herrmann

8.3 The empty atom

Subject:

Since the atomic nucleus is small and heavy compared with the electron shell of an atom, it is often concluded, that the greatest part of the atom is empty. The electrons are supposed to be point-like. (Sometimes it is said explicitly, sometimes insinuated). Thus from the space occupied by the whole atom, only a very tiny fraction is occupied by matter: “An atom essentially consists of empty space, populated only by a minuscule nucleus and the electrons.”

Deficiencies:

1. The claim about the emptiness of the atom depends on the model that is applied. It is true if we imagine or model the electron as a small individual with the remarkable property of being able to move without having a trajectory. To describe such a behavior the concept of probability density was introduced. According to another model (the substance model) the electron occupies the whole space that is covered by its wave function. The square of the wave function is a measure for a kind of electron matter distributed in space. Hence, the size of the electron is that of its orbital. Since the orbital has no well-defined boundary surface, Bohr’s radius could be taken as an effective size of the electron.

2. If one has opted for the point-like electron model, it would be consequent to apply the same model to the nucleus, i.e. the protons and neutrons which consist of point-like quarks. In this case not only the greatest part of the atom would consist of empty space, but the whole atom, and thus the whole world. Obviously this statement is rather useless.

Origin:

Rutherford’s experiment which suggested that the nucleus is a small compact body.

Disposal:

The original intention was to express a simple fact: The mass of the nucleus is much greater than that of the shell. However, this can be said without referring to the problematic empty space. By the way, one should not forget that the shell does so poorly only when comparing masses. But mass is only one of the extensive quantities that characterize a particle. Regarding the electric charge, the nucleus and the shell are at par. The same is true for the angular momentum. And when comparing the magnetic moments, it is the shell that wins.

Friedrich Herrmann

8.4 Electronic shells

Subject:

To explain various properties of the atom, as for instance the periodicity of atomic radii or ionization energies with increasing atomic number, one makes use of the shell model. In order to substantiate the existence of shells, one often represents $r^2 \cdot \rho(r)$, i.e. the electron density of a many electron atom, multiplied by r^2 as a function of the distance r from the nucleus. The corresponding graphical representation also shows, so it is said, "that the probability of finding the electron in the region occupied by the nucleus is extremely small".

Deficiencies:

Whereas the electron probability density decreases monotonically as r increases, Fig. 1a, the function $r^2 \cdot \rho(r)$ is zero at the center, i.e. in the region of the nucleus and has several maxima for increasing r. Finally for great values of r it tends to zero again, Fig. 1b. The function of Fig. 1b is not the normal spatial probability density, but the probability per radius interval dr. Some textbooks point out that a trick is used, others do not. Anyway, it is hardly avoidable that the reader mistakes the expression corresponding to the vertical axis for the density itself. Our experience with physics students

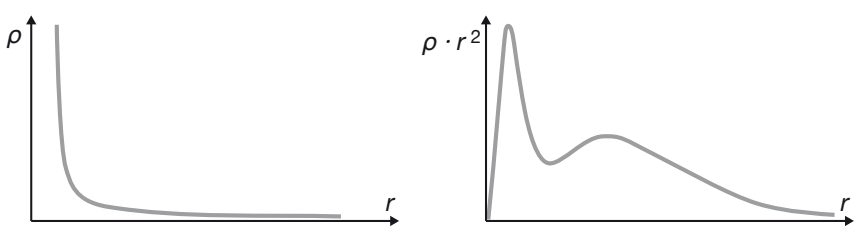


Fig. 1. (a). Probability density in an atom. (b) Density integrated over solid angle

at the university showed us that the students memorize the shape of the diagram and that the vertical axis represents the probability density. In particular they believe that the probability density is zero at the center and that there are shell-shaped regions where the density is particularly high.

The following example shows that a representation of $r^2 \cdot \rho(r)$ can indeed disconcert when trying to get an idea about a density distribution. We ask for the mass distribution of a massive glass sphere. We plot both its mass density $\rho(r)$, Fig. 2a, and the expression $r^2 \cdot \rho(r)$, Fig. 2b as a function of r (the distance from the center). Obviously, in order to get an idea about the mass distribution in the sphere, it is better to look at Fig. 2a.

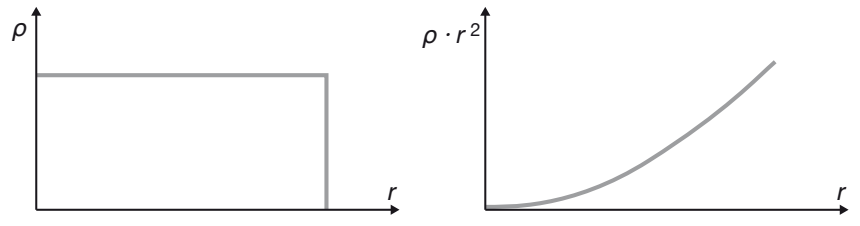


Fig. 2. (a). Mass density in a massive spherical object. (b) Mass density integrated over solid angle

The claim that it is much less probable to find an electron within the nucleus than further outwards, is of the same kind as the following statement: It is much less probable to find a winner of a lottery prize in Berkeley than in Nevada. In statistics one refers to this error as biased sampling.

Origin:

Apparently some physicists are not satisfied with disposing of a convenient model which makes some correct predictions. They seek to recognize the shells in the density distribution. Moreover, they seem to have problems with the idea that the probability of finding an electron inside the nucleus is not zero.

Disposal:

The representation of $r^2 \cdot \rho(r)$ does not have a substantial advantage, but is the cause of misconceptions. We recommend to represent only the density $\rho(r)$.

Friedrich Herrmann

8.5 The wave function

Subject:

"The wave function itself has no direct physical meaning."

"… ψ does not represent a quantity that can be directly measured like a length or field strength."

"The fact that the wave function is not real, but complex, reflects, among other things, that $\psi(\mathbf{r},t)$ has no real physical meaning like for instance the electric field strength $\mathbf{E}(\mathbf{r},t)$ of a light wave in classical optics or electrodynamics respectively (in quantum electrodynamics also $\mathbf{E}$ does not have a real physical meaning)."

Deficiencies:

There are tenets in physics, that appeared to me particularly transcendental, when I still was a student. They seemed to be propositions which had to do with the very foundation of science. They were unexpected and I could not really understand them, and it was not clear to me for what they were needed. Among these propositions is the one that says that the wave function cannot be measured directly. Although it was mentioned by the professor only casually, the claim was engraved in my mind. And it has found its way into the school books.

So, why can the wave function not be measured directly? Two kinds of justifications can be found: 1. The fact that it is a complex quantity. However, there are other complex quantities. Everyone knows how to deal with them. No warning that these are quantities that are not measurable. 2. The claim that the absolute value of ψ reflects the electron density distribution, but that the phase is arbitrary and unmeasurable. But this is not quite correct. The phase manifests in the current density (often called the probability current density) and this can be measured. Thus (at least for one-particle wave functions) the wave function is completely determined by the density and the current density.

When emphasizing that a quantity cannot be measured directly, one should specify what is meant by a direct measurement. In this context one often mentions the electric field strength, but the measurement of this quantity is not what one really would like to call a direct measurement. One uses a test charge that modifies the field that is to be measured in such a way that, at the position of the test charge, there is not the slightest resemblance with the original field.

Origin:

Max Born's probability interpretation hinders us to associate any intuitive idea with the wave function.

Disposal:

The claim that the ψ function cannot be directly measured does not make sense as long as it is not specified what is meant by "directly" measured. The want to spell such a warning will probably disappear when using another model (or "interpretation") of the square of the wave function, for instance that of Schrödinger and Madelung who interpret ψ squared as a measure of a continuous mass and charge distribution of the electronic cloud.

Friedrich Herrmann

8.6 Indistinguishable particles

Subject:

When treating the laws of quantum statistics, it is emphasized that particles are identical or indistinguishable.

"Two particles are called identical, when the result of the measurement of any magnitude or observable of the system is invariant with regard to an interchange of the particles."

"Two particles are called identical, if they coincide in all their intrinsic properties (mass, spin, electric charge etc.): There is no experiment which allows to distinguish between the particles. Thus, all the electrons of the universe are identical, just as all the protons or all the hydrogen atoms."

Deficiencies:

As a student I had always a feeling of uneasiness when the indistinguishability was mentioned: Is the statement a triviality or does is concern one of the strange properties of the quantum world that are difficult to understand? That the statement gets into our mind only with difficulty has probably two causes.

1. The two particles that are supposedly indistinguishable, can well be distinguished. Imagine two electrons: one at position r_L (left), the other at position r_R (right). Surely the are similar in many respects: the same mass, the same electric charge, the same spin, the same state of excitation, and whatever properties they may still have. (One says, they coincide in their intrinsic properties.) But there is a trait in which they are different: position. One is located at r_L, the other at r_R. They thus can be distinguished.

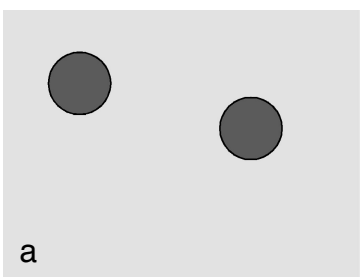


Fig. 1. Exchanging two circles on the computer screen

Actually, what matters in statistical physics is not the indistinguishability of particles but that of states. We again consider our two electrons. We consider a state, in which one of them is located at r_L and the other at r_R. We now bring the left particle to r_R and the right one to r_L. The state we have obtained cannot be distinguished from the previous state. Both states are identical – even when using the word "identical" in its colloquial meaning. However, in each of the two states, before or after the exchange, the two particles can be distinguished: the one is at the left, the other at the right.

The situation is similar to the one in the following "experiment": With a drawing program we create on the computer screen two full circles with the same radius and the same color, Fig. 1a. Now we move both circles on the screen with the mouse, Fig. 1 b, and finally reconstitute the old image, Fig. 1c. Now, when moving the circles we had exchanged their position. The images of Fig. 1a and Fig. 1c are indistinguishable though. However, both circles in one image can be distinguished.

2. Can it be that by interchanging two particles truly the same state results? Is there really no means to see that the "new" state has arisen from the "old" one by an interchange? We would not have this problem if we had not, by a long exercise in classical mechanics, adopted the practice to imagine a particle as a small being, which is characterized, apart from the values of certain physical quantities, by something that could perhaps be called its soul. Instead of trying to get rid of this habit when doing quantum physics, the idea is nurtured and cherished by the very language of quantum physics. It is interesting to note that we have the problem when electrons are concerned. We do not have it when interchanging two holes in a semiconductor. We do not imagine a hole as an individual in the same sense as an electron. It seems that they correspond much better than the electrons to the two circles on the computer screen.

Origin:

Classical mechanics has to do with individual bodies. The concept of an individual is appropriate if the corresponding system has properties which it retains and which allow us to recognize it at other instants of time and in other environments. In our every-day experience the properties that characterize an individual are mainly the shape and the distribution of the substances that constitute the system. If however, the number of degrees of freedom becomes very small until finally there remain only mass, momentum, angular momentum and position, the concept of individual simply melts away. In other words: The concept becomes justified and unambiguous asymptotically for systems with a great number of degrees of freedom. For this reason, the concept of an individual particle of classical mechanics is not a good basic concept for quantum mechanics.

Disposal:

In our everyday language we do not have a problem in describing objects that are not individuals in the sense of classical mechanics. Examples are a cloud in the sky or the flame of a candle. Is the cloud five minutes later the same cloud as that five minutes ago? Is the flame after 5 seconds still the same flame? The answer of an unbiased person to these questions may be no more than a shrug of the shoulders. There simply is no problem.

In quantum mechanics much would be done if the language and the models that underlie it would be slightly modified. So, the electron could be introduced as an indivisible portion of a substance with a well-determined mass, electric charge and angular momentum. If now we have one such portion on the right hand side and one on the left, and we interchange them, it is not difficult to see that the final state is the same as the initial state. The expectance of something like a soul will not come up from the very beginning.

Friedrich Herrmann

8.7 Photons and phonons

Subject:

Physics text books for the upper secondary school introduce the concept of a photon: either as an energy portion that is exchanged in the process of absorption or emission of light, or as the constituent particles of light. Phonons on the contrary are not mentioned in the majority of these books. This comes along with the fact that physics students at the university have a rather concrete idea of photons and a rather pale idea of phonons.

Deficiencies:

There is a far reaching analogy between photons and phonons. The classical theories of light and of sound have much in common, just as the corresponding quantum theories [1, 2]. The analogy manifests itself in various effects.

An example is heat transport with the one or the other particle. The carrier particles of a heat transport in a heat conductor of a material that is not an electric conductor are phonons. (In an electric conductor electrons dominate the process.) The process is diffusive, i. e. there is a continuous production and annihilation of phonons. Very similar is the heat transport within the sun from the reaction zone outwards. Here, the carrier particles are photons that are steadily emitted and absorbed.

The analogy also shows up in the temperature dependance of the energy of the phonon and the photon system in thermodynamical equilibrium. In both cases the energy varies as the forth power of the temperature (which in the case of photons is known as the Stefan-Boltzmann law).

Thus both kinds of particles have much in common and do not merit to play such a different role in the teaching of physics.

Origin:

Phonons entered the physical scenery via the quantum-physical treatment of lattice vibrations. On the contrary, photons as the particles of light have a century-old tradition. In addition, single photons can easily be detected. Detectors for gamma and X ray photons exist since a long time, but today photons of visible light can also be detected with material that is not too expensive.

The fact that phonons are often called “quasi particles” may contribute to the belief that the phonon is a more abstract concept than the photon. Quasi particles are particles which owe their existence and their properties to their local environment. Actually it looks as if the distinction between quasi particles and the so-called normal particles is becoming obsolete, since we just learn that the normal particles owe their properties to the Higgs field.

Disposal:

1. Less reticence to introduce phonons and to treat them as particles. They are no more difficult than photons. We have found text books for the school which introduce gluons. So why not phonons that are certainly nearer to everyday physical phenomena than gluons.

2. A little more prudence when introducing photons.

3. More reticence in using designations like “quasi” or “virtual”. Such terms create uneasiness in the students’ minds about a concept and hardly explain anything.

[1] *Ashcroft, N. W., Mermin, N. D.*: Solid State Physics, Holt, Rinehart and Winston, Inc., Orlando (1976), p. 453: “In that theory [quantum theory of the electromagnetic field] the allowed energies of a normal mode of the radiation field in a cavity are given by $\left(n+\frac{1}{2}\right)\hbar\omega$, where ω is the angular frequency of the mode. It is univesal practice, however, to speak not of the quantum number of excitation of the mode, *n*, but of the number, *n*, of *photons* of that type that are present. In precisely the same way, instead of saying that the normal mode [in a crystal] of branch *s* with the wave vector **k** is in its $n_{\mathbf{k}s}$ excited state, one says that there are $n_{\mathbf{k}s}$ phonons of types with wave vector **k** present in the crystal.”

[2] *Vogel, H.*: Gerthsen-Kneser-Vogel, Physik, Springer-Verlag Berlin (1977), p. 598: “A lattice vibration with the angular frequency ω can, just as the oscillation of a single particle, only have energy values whose differences are entire multiples of $\hbar\omega$. For this reason a light wave for instance can exchange only an entire multiple of this value with the crystal lattice. With the same justification as in the case of the electromagnetic wave field this is interpreted as the existence of *acoustic quanta* or *phonons* of energy $\hbar\omega$.”

Friedrich Herrmann

8.8 Photons in the sun

Subject:

"A photon, that is a 'particle of light' that is formed during the fusion processes in the core of the Sun, moves at the speed of light, i.e. at 300,000 km/s – but only until it hits a particle and is scattered from there in a different direction. Inside the Sun, matter is extremely densely packed, so a photon can not move far in one direction without being redirected – often just fractions of a millimeter. Outwardly, this distance gradually becomes slightly longer. To calculate how long a photon takes to reach the surface through random scattering from inside the sun, one has to make some assumptions about the structure of the sun, for example about its exact density distribution. One obtains values between 10,000 and 170,000 years."

"You can calculate how long it would take one photon to 'diffuse' by scattering through the core to the bottom of the convection zone, and this has been done (it's about 170 000 years)."

Deficiencies:

One clings to the photon, the particle of light, the small being, the small body, or the wavelet that rushes or wobbles through the world. Even with the electron, the idea that it is a small individual causes some difficulty of understanding. One has to make incomprehensible additional assertions ("electrons are indistinguishable") in order to be able to maintain the language that one uses (and thus the mental model of it). The photon is even worse. Although Einstein's sentence is famous and well-known, apparently it is not taken seriously and dismissed as a bon mot of the great master [1]. Note that his statement does not come from the early days of the photons. Quantum electrodynamics was born long ago.

But what is the problem with our quotes (that are rather typical)? That they not only suggest, but clearly say that a photon from the reaction zone inside the sun reaches the surface. "A photon" means in accordance with our normal way of speaking: A photon starts its journey and the same photon, the same individual, arrives near the surface of the sun after 100,000 years. Nothing is left from the warnings and the reservations of quantum electrodynamicists. One would like the photon to be a small creature, and so one makes it a small creature.

The sentences should be doubted, alone for the reason that the number of photons arriving at the sun's surface is about 3000 times that of the photons that started the journey.

By the way, no one seems to come up with the idea of describing thermal conduction in a copper rod in a corresponding manner: one would say that a phonon moving at the speed of sound would take a minute to cross the 30 cm long heat conductor.

Origin:

The photoelectric effect seems to show that light consists of particles, and we imagine a particle as an individual. This approach is seductively catchy. All warnings, neither that of Einstein nor those of quantum electrodynamics, seem to fall on deaf ears.

We still carry with us the shackles of the mechanistic worldview that served us so well until the end of the 19th century. In a more general form, we come across the conception as an attitude that philosophers of science call reductionism. It is believed that the description of the world becomes simpler, or that one understands better how the good Lord has made the world, by describing the perceptible phenomena by atoms, the atoms by protons, neutrons, electrons, etc., the protons and Neutrons by quarks, etc. ad infinitum.

Every now and then one asserts that this is just a model, the particle model, but that is probably just lip service. Because if you were convinced that it was just a model, you would speak otherwise of the photons. A model always means: "It's like …". For example, like a small body. In fact, there are situations, processes or states in which the radiation behaves like small bodies, even if they are not too frequent.

The idea that the world is made up of small individuals seems to be reassuring: the microworld is like the macroworld, like the familiar world around us.

Many a physicist also seems to have lost the understanding of what one can expect from a beginner, one who wants to learn physics. It seems to be assumed that the small swarming individuals are the only thing acceptable for the explanation of the physical world. In fact, the unindoctrinated learners do not have the problem that the physicist presumes: they have no problem talking about a cloud moving in the sky, and of course they find it natural that, if the cloud's motion is followed long enough, the original cloud no longer exists, and instead another one has formed. They have no problem with the term individual and the indistinguishability. This is first suggested to him by a certain teaching tradition.

Disposal:

To describe the energy transport within the sun, one does not necessarily need quantum electrodynamics. But at least one should say that absorption and emission processes take place, so that the idea of individuals swarming around does not arise. Point out that the same process takes place in the troposphere and contributes to the cooling of the Earth, only translated into the infrared.

It should also be pointed out that the usual heat conduction is of a very similar kind, only with phonons instead of photons. One learns two things at once: something about the photons and something about the phonons.

For many purposes, it would be enough to say that a temperature disturbance in the interior of the sun takes 100,000 years to show up on the surface of the sun. Or one says it more exactly: The energy transport happens with electromagnetic radiation. This radiation is almost in thermodynamic equilibrium, i.e. it is blackbody radiation; its temperature decreases from the inside to the outside; the temperature gradient is very small, but it is this gradient that causes the flow; the transport is dissipative; on the way the entropy increases by a factor of 3000 (equal to the ratio of the temperatures).

If you want to express it in photons: they are emitted and absorbed and new ones are emitted, and so on. But do not be deluded. To say that they move from one atom to another is reckless.

And you can also make a little epistemological remark: Explain the nature of a theory. A theory is a mathematical description of the world. A theory is not wrong or right, but only more or less suitable for a given purpose.

Friedrich Herrmann

[1] Albert Einstein wrote in 1951 in a letter to his friend Michele Besso:
"All these 50 years of conscious musing did not bring me any nearer to the answer of the question 'What are the quanta of light'. Today every bounder believes to know it, but he is wrong…".

8.9 The particle model of matter

Subject:

At school, the particle model of matter is introduced. It can be found in textbooks of the elementary, intermediate and advanced levels. It seems to be an important topic. What is meant by the particle model? Here are some statements that are highlighted in the books:

Textbook, ages 10 to 12:

> The particle model
>
> 1. All substances consist of particles (small spheres).
> 2. The particles are in constant motion.
> 3. Forces occur between the particles.

Textbook, ages 13 to 14:

> Model for gaseous bodies: Gases consist of particles that move freely in space.

Textbook, upper secondary school:

> Model representation of ideal gases:
>
> 1. In collisions the particles behave fully elastically.
>
> 2. Except during the collision, the particles exert no forces on each other.
>
> 3. The particles are elastically reflected like spheres on the walls of the container.
>
> 4. In the disordered movement of the particles all directions of movement are equal.
>
> 5. The intrinsic volume of all particles taken together is negligible relative to the volume of the gas.

Deficiencies:

First two explanations: What is meant by a particle and what is meant by a model?

Particle: In the colloquial, and also in the scientific and technical sense: a small object. In general there are many of them. Typical examples are a dust particle or a soot particle. (In contrast to the non-diminutive „part“: a part is not an object, but just a part of something, of an object.)

Model: A model is always a model of something else. Suppose B is a model of A. A consists of elements between which certain relationships exist. Since B is a model of A, B must also consist of elements that are linked by relationships. The elements and relationships of A are mapped on those of B; one can set up a kind of translation table. One can now draw conclusions in B and translate them into inferences within A using the translation table. If such inferences are often correct in A, the model is a good model; if they are often wrong, the model is bad. In any case, in most properties the original and the model do not match. There are no wrong models and correct models, but only more or less useful models.

Now for our quotes. It is unclear why the term “model” is used. Who is a model of whom? Should particles be the model of atoms and molecules? Then one would have to explain why an atom **is** not a particle. Moreover, the texts consistently say that the atoms and molecules themselves are particles.

An appropriate use of the term model can be found for example in Bohr’s model of the atom. The atom (A) is built and behaves in some ways like a planetary system (B). In most properties, the atom and the planetary system are not at all similar, but in some that are important in a certain context, they are.

Only in a chemistry book I found that the author was trying to justify the term model, but in a way that I did not fully understand:

> “However, the smallest particles are not visible without aids. ...
>
> This model is therefore a thinking aid. It is a thought model about the possible structure of the substances.
>
> When using the particle model one imagines that the particles of the substances are very similar to small spheres. …”

Now invisibility is certainly not a reason to speak of a model. Not seeing the air does not make us introduce a visible model of the air. And where is the thinking aid? Should one believe, for example, that “in reality” the substances do not consist of atoms at all? The atoms are only a mental aid?

Actually, to speak of a particle „model“ has a meaning, namely, when the objects we are talking about no longer have the essential characteristics of the colloquial particles: when at very low temperatures the uncertainty of the position of the atoms becomes significantly larger than their diameter, or if two (or more) „particles“ are entangled, so that one can only speak of a delocalized particle, or if the particles have so few internal degrees of freedom that two „particles“ are no longer distinguishable, so that after a permutation the particles are in the same state as before.

Now these are states and processes that one certainly does not have in mind when introducing the “particle model”. By the way, when these phenomena, which challenge somewhat the idea of a particle, are finally treated, the term particle is often used with a surprising unconcern.

But is it really bad when sometimes a word does not quite fit? The problem is that our textbooks contain many phrases that suggest that something profound is discussed. It is one of many trifles which, taken together, make physics so unsightly; that make physics seem more complicated than it is. What arrives at the students is: It is not important to understand; it is important to repeat the expected words in the exam.

Origin:

The term particle model probably comes from the curricula. I can not say how it got in there. It is no wonder that the textbook authors are a bit helpless. They have to write something about it, but do not know what. Or perhaps they believe that the term particle model sounds so pretty, so profoundly epistemological?

Disposal:

We, teachers, curriculum makers, textbook authors are responsible for the fact that physics is the most hated school subject. What is needed is disarmament. The disposal in our specific case is simple: Leave the term model out. It is appropriate anyway that the students first learn the physics. If you have time left in the upper secondary school, then you can also discuss some metaphysics.

Friedrich Herrmann

8.10 Wave-particle duality

Subject:

As a reminder the Wikipedia definition, first from the English article, and second from the German (translated into English):

> Wave–particle duality is the concept in quantum mechanics that every particle or quantum entity may be described as either a particle or a wave. It expresses the inability of the classical concepts "particle" or "wave" to fully describe the behavior of quantum-scale objects.
>
> Wave-particle dualism is an insight of quantum physics, according to which the properties of classical waves as well as those of classical particles must be attributed to the objects of quantum physics. Classical waves propagate in space. They weaken or strengthen each other through superposition and can be present at different locations at the same time and thereby have different effects. A classical particle can only be present at a certain position at a certain instant of time. Both properties seem to exclude each other.

Deficiencies:

Already as a student I felt uneasy when there was talk of the wave-particle duality. What was meant? Should something be explained or just named or even veiled?

1.The behavior of an electron, photon or other "quantum entity" is presented at first as contradictory. But then one learns that there is no real contradiction, because there is a duality. Did you understand? Maybe not quite, but you know what to say in the exam.

The problem is that apparently it is presumed that only one of two mutually exclusive models can be taken into consideration.

- Either the electron (or the photon…) is a "particle". But what is a particle? A small body, a small individual whose location is described by the coordinates of a single point. This is not simply the position of the center of mass; it is the "position" of the whole particle. So there is no other choice than to imagine the particle as point-like.
- Or the electron is a wave. The normal idea of a wave is somewhat like this: You have some kind of wave carrier, (for example water or air) and on or in this carrier the wave is travelling: a change of state, which is propagating, and which is imagined, according to its name, to be wavelike, i.e. there is some up and down, or big and small; somewhat periodic, but not exactly periodic. A wave has an extension, both lengthwise and crosswise (except if it is a wave in a rope, or a surface wave). That the wave has a spatial extension is self-evident, but the point-mechanically socialized physicist seems to consider it necessary to emphasize that a wave is not point-like, see above: Waves "…can be present at different places at the same time." Do we really need to explain this?

Since both models don't really seem to fit, something needs to be said. And one actually says something: there is a duality.

The learner is left with a feeling of frustration, because the magic word duality does not explain anything. It is just a euphemism for expressing that something is logically inconsistent.

However, the problem would not have arisen in the first place if one had not started by presenting the electron or photon as a small individual.

One might believe that there is no other choice; particles and waves are the only categories of human thought that come into question in our case. I mean, they are rather the thinking categories of the physicist.

For someone who has no training in physics, there are other models available, that can be applied to what physics in its need calls a quantum entity.

Is it really so difficult to imagine an object that is not represented by a single spatial coordinate but by a distribution? Everybody knows things that are said to be located somewhere, but one does not demand that the thing is the same "individual" now and a little later. Think of something like a cloud or a flame, or even better a hump of a wave on the water. It is somewhere, but not in one point, it is extended, and two of them even show interference. Would anyone here speak of duality?

2. It is often said that the electron (or photon) *sometimes* behaves like a particle and *sometimes* like a wave. Such statements are probably based on the fact that one speaks of the property of the electron only when one is making a measurement: Either one looks at the interference pattern, or at the pixels of the detector. Doesn't one see the double nature here in all clarity? Not really. Because one must ask: the nature of which electron at which time? The interference pattern results from the wave function of the electrons before they were "detected". The blackened pixels afterwards. So the two properties refer to electrons in different states: once in a state with a sharp momentum and unsharp position, and once with a sharp position and unsharp momentum. One concludes from this that electrons sometimes behave like this and sometimes like that. It would be clearer to say that electrons can be in different states, and actually in an infinite number of different states. Among these there are two types of extreme states: In one of them they have a sharp position, in the other a sharp momentum.

3. I can't resist analyzing the Wikipedia definition somewhat linguistically: (The language in Wikipedia is a collaborative work. So there must be some consensus about the result.) It expresses, in my opinion, the helplessness in this context. Let us begin with the generic term of "duality". In the German version, duality is an "insight" (Erkenntnis). Since I would not have expressed it in this way, I switched to a few other languages in order to see if there dualism is also an insight. But no, it is not. In English it is a "concept", in French and Dutch a "principle" (principe, beginsel), in Spanish a "phenomenon" (fenómeno), in Chinese a behaviour (行为). Of these I prefer the Chinese one. Nowhere is it a "property", by the way.

It is also interesting to see how philosophers, theologians etc. deal with the term duality. That alone should make us suspicious.

But let us go on reading: Quantum objects have "equally" the properties of particle and wave. Equal is not the same as simultaneous. Again, the formulations in other languages are interesting. Let's look at the French one, for example: according to it, the particles sometimes show particle and sometimes wave properties, i.e. not simultaneously.

And finally, the little word that is so often used in physics lessons to disguise things: Particle and wave properties are "attributed" to quantum objects. They do not *have* the properties, but the properties are attributed to them. Where else do we use the word attribute? I recommend (as on other occasions) to consult the website of *Linguee*.

Origin:

There were the two categories at the end of the 19th century: light was an ether wave; the recently discovered electrons were small corpuscles.

When it became clear that electrons also show interference, and light is quantized, the idea arose that both light and electrons have a strange dual nature.

It was not until shortly afterwards that the puzzle was solved: in 1926 came the Schrödinger equation and in 1927 the uncertainty principle.

This succession probably determined the teaching pattern.

If the wave function had been at the beginning, the idea of duality would probably not have arisen.

Disposal:

Avoid telling something that will hinder you later, so

- do not use the concept and the word duality or dualism;
- do not raise the expectation that the particles are point-like;
- do not associate the ability to interfere with the sine wave.

We imagine the electron as a thing with the charge e and the mass m_e. Sometimes it is larger, sometimes smaller, sometimes more "monochromatic", sometimes less, but always capable of interference.

The quantity psi as a function of position and time – the solution of the Schrödinger equation – contains everything there is to say (at least as long as we are not yet dealing with particle physics and the Standard Model). Nor does it hide anything that needs to be interpreted or put into mysterious words.

Friedrich Herrmann

8.11 Quanta and quantization

Subject:

Quantum theory, as its name suggests, is a theory about quanta. But what are quanta? What is a quantum? We are concerned here only with the use of the word. Here are some examples:

1. “In physics, quantum is understood as an object produced by a change of state in a system with discrete values of a physical quantity. Quantized quantities are described in the framework of quantum mechanics and subfields of theoretical physics inspired by it, such as quantum electrodynamics. Quanta can occur only in certain portions of this physical quantity, they are consequently the quantization of these quantities.”

2. “A phonon is the elementary excitation (quantum) of the elastic field.”

3. „In physics, a quantum … is the minimum amount of any physical entity (physical property) involved in an interaction. The fundamental notion that a physical property can be ‘quantized’ is referred to as ‘the hypothesis of quantization’. This means that the magnitude of the physical property can take on only discrete values consisting of integer multiples of one quantum. For example, a photon is a single quantum of light (or of any other form of electromagnetic radiation).“

4. „Quanta: The particles obtained by the complementary approach to the wave fields. In particular, one understands by it the light quanta, the particles which are to be assigned to the electromagnetic field. According to the quantum theory of the fields, each field has its quanta; thus, to the nuclear field belong the mesons and to the matter field, which causes the chemical forces, the electrons.“

Deficiencies:

What is the meaning of a word? What concept does it designate? This is decided solely by the way it is used. This means, especially in the context of colloquial language, that a word can have several or even many meanings.

In physics, and especially when a word appears to be a technical term, one would want the meaning to be unambiguous. Often, however, this is not the case even in physics. The meaning is then (hopefully) revealed by the context. In mechanics, for example, “force” is usually the name for the quantity F, but sometimes, especially in word combinations, energy is meant. With “current” sometimes a phenomenon is meant, sometimes the quantity electric current strength and sometimes the electric charge.

A particularly dazzling word in this context is the term “quantum”. There are innumerable word combinations with the word quantum: Quantum condition, quantum property, quantum hypothesis, quantum tunneling, quantum object, quantum number, quantum statistics, quantum interference. As an undergraduate, I learned quantum physics with a textbook titled “Quanta”. With the advent of quantum computers, there has been a veritable inflation of the word quantum.

Let’s go once again through the quotes:

1. Here, a quantum is sometimes an object, sometimes a portion of a physical quantity.

2. A quantum is an excitation.

3. A quantum is the minimum amount of a physical entity or a physical property. But also a particle, namely the photon, can be a quantum.

4. Also electrons and mesons are quanta.

It is obvious that the term is a generic term – but for what?

Apparently for several classes of terms: namely on the one hand for particles and on the other hand for elementary portions of physical quantities. Nevertheless, we cannot make a definition out of it, because

- not every particle, not even every elementary particle is called a quantum;
- for the energy there is no universal elementary portion, but one would like to speak of energy quanta, for instance in connection with the harmonic oscillator; and for the entropy there is an elementary portion, namely k_B, but one would not like to say that the entropy is quantized.

Origin:

The linguistic usage was not quite clear from the beginning. In his famous work of 1905 Einstein uses both the designation “energy quanta” and “light quanta”. It was not yet clear that here it goes back and forth between a statement about the values of a physical quantity (energy) and one about what later was called particle (photon). However, as more and more “elementary” particles were discovered and more and more quantities turned out to have values that are integer multiples of an elementary amount, it should have become clear that one should distinguish conceptually between object (particle) and quantity (energy). But this differentiation has not taken place.

Disposal:

There is no institution that prescribes how words are to be used, not even the so-called technical terms. I do not want to give recommendations to the physicists for the use of terms. If I make recommendations, they are directed only to teachers, at school and the university.

My first recommendation would be not to use the term quantum as a name of particles at all. The name simplifies nothing, it explains nothing. Do not call the particles of the light quanta, but photons. And do not call photons quanta of energy (as they are also not quanta of momentum or angular momentum).

My second recommendation would be not to use the word for the elementary amounts of physical quantities either.

On the other hand, in my opinion, the verb quantize is useful. Thus, the fact that the values of certain physical quantities (in a closed system) are integer multiples of an elementary value can be expressed briefly and clearly: electric charge, angular momentum, magnetic flux... are quantized. If one uses the word in this way, however, one should also say that the quantity of substance is quantized.

The elementary quantum

$$\tau = 1.66 \cdot 10^{-24}\ \text{mol}$$

is just the reciprocal of the Avogadro constant. The reason that it is usually not expressed in this way is probably because this quantization had been discovered long before the emergence of the quantum theory.

However, it would be rather clumsy to speak of the quantization of the energy. Like other quantities, the energy of a given system usually takes on discrete values. But shall we call this quantization?

Alternatively, one could dispense with the word altogether in this context. One would not lose much and perhaps gain some clarity.

I am afraid, however, that in view of the proliferation of the quanta, such recommendations are unrealistic. But perhaps one can at least give the following advice: Moderation in dealing with the word.

Friedrich Herrmann

8.12 Degeneracy

Subject:

"In quantum mechanics, an energy level is degenerate if it corresponds to two or more different measurable states of a quantum system."

"Degenerate matter is a highly dense state of fermionic matter in which the Pauli exclusion principle exerts significant pressure in addition to, or in lieu of thermal pressure… Degenerate matter is usually modeled as an ideal Fermi gas, an ensemble of non-interacting fermions."

Deficiencies:

I confess that as a student I had a problem with the term degeneracy or degeneration for a long time. Is both times the same thing meant and I just didn't understand it? Not that I could not have explained it in the exam. My answer would just have been different depending on the context.

I just recently learned that students today still have the same problem.

And another observation: The term "degenerate" is certainly not an appropriate expression. Degenerate is something that is not as it should be or as one would like it to be. But why are states degenerate if two eigenvalues are equal for symmetry reasons? Likewise the Fermi gas. Why should the ideal gas be the ultimate reference? All around us there are Fermi gases. With such a name, one certainly expects something that deviates more from the normal.

And finally: Is it necessary to introduce two technical terms for one and the same thing: Fermi gas and degenerate gas? Let's assume that in order to be able to talk about physics, one has to know the names of 3000 terms. If one introduces two terms for each concept, this makes 6000 words to be learned.

Origin:

The progress of physics is a process of evolution. There is no foresight, or almost none. And there is no institution that ensures that the language becomes coherent or that superfluous terms are thrown out.

Disposal:

At the very least, make students aware that you use the word in two different meanings in the same lecture. Here I would like to praise Wikipedia (German version): Under the keyword "Degenerate matter" one finds the sentence: "Here the term *degeneracy* has a different meaning than in the case of degenerate energy levels."

But perhaps one can also decide to do without at least one of the two uses. (This is what Tipler does. There is only a Fermi electron gas).

Of course, I know the argument: students need to be enabled to understand other texts. Certainly, that may be one of the learning goals of the lecture. But a more important one is to make students understand the subject matter, and technical word proliferation is a serious obstacle to that. One of the consequences is the notorious unpopularity of physics as a subject at school.

Friedrich Herrmann

8.13 The shape of photons

Subject:

We are interested in how photons are talked about and what ideas are conveyed about their size and about their shape.

Let's look into Wikipedia, first into the German version:

> "**Photons** (from Greek φῶς *phōs* 'light'; singular 'the photon', accent on the first syllable), also **light quanta** or **light particles**, are, in illustrative terms: the energy 'packets' that make up electromagnetic radiation.
>
> Physically, the photon is considered as an exchange particle. According to quantum electrodynamics, as a mediator of the electromagnetic interaction, it belongs to the gauge bosons and is thus an elementary particle. The photon has no mass, but energy and momentum – both proportional to its frequency – as well as angular momentum."

or into the English Wikipedia:

> „A **photon** (from Ancient Greek φῶς*,* φωτός *(phôs, phōtós)* 'light') is an elementary particle that is a quantum of the electromagnetic field, including electromagnetic radiation such as light and radio waves, and the force carrier for the electromagnetic force. Photons are massless, so they always move at the speed of light in vacuum, 299792458 m/s (or about 186,282 mi/s). The photon belongs to the class of bosons."

and finally into some school textbooks:

> "During the interaction between light and matter, the energy is always transferred in small portions. These portions are called light quanta or photons."
>
> "These portions of energy are called photons. We say: Light energy is quantized....
>
> Proposition: The energy of electromagnetic radiation with frequency *f* is effective in quanta $W = h \cdot f$, called photons. Light energy is quantized."

Deficiencies:

Before I get to the actual topic, namely how to talk about photons, briefly to the simpler question of what a photon is, and also what it is not.

What is understood by a photon in physics is clearly stated in one of our Wikipedia quotations, namely the English one: it is an elementary particle, it is the quantum of the electromagnetic field, it is the particle which mediates the electromagnetic interaction. According to general linguistic usage in physics it is not a portion of energy, even if this is said in the two cited school textbooks, and even if the German Wikipedia offers it as an illustration. Because if somebody claims that a photon is an energy portion, then immediately the question arises, why he does not call it a momentum or angular momentum portion, and why then an electron should not also be called an energy portion.

I do understand that textbook authors want to avoid the ugly questions: Where is the photon located? How big is the photon? Which way does the photon take? But the recourse to the energy portion is not a solution.

Now to our question about size and shape. Nothing is said about this in any of the quotations. Is this perhaps because we have asked a bad question? As is well known, there are bad questions, for example the question about the color of an electron. Is the question about the shape of the photons such a question?

However, if it were so, one would not be allowed to make some statements which are usually made quite unconcernedly in the context of photons.

If one says that the photon passes with a certain probability through one slit and with the same probability through the other slit, one assumes that the width of the photon is smaller than the slit width. If one says, the photons in the interior of the sun reach at most one millimeter before they are absorbed, one assumes that they are not longer than 1 mm. But why does one not pronounce this clearly? Might this be caused by the idea that the photon is point-like? Of course, one does not dare to say something like that.

Since if the energy of the photon is *h* times *f*, it must have a frequency and also a wavelength, and since one gives a single sharp value for the frequency, the photon should be infinitely long. So one seems to be in trouble.

The problem is that if we don't say anything about it, the learners will make up their own minds. Do we, the teachers, really want that?

Of course, every physicist knows the reason why one does not like to answer the question. We are dealing with a particle, with an entity of the real world, which exists, but which is not "localizable", or maybe only sometimes and only a little bit, and that it sometimes has to share its identity with another particle.

Concepts like these do not seem to be unacceptable at all for some people. Compare for example with what is sometimes said about the soul, which exists somehow in space. It is only weakly localized and its trajectories are not well defined. Of course, we prefer not to orientate ourselves on such fantasies. After all, we have a coherent theory which predicts which results will give our measurements with which probabilities. And it is about nothing else than to find a language which allows to communicate the content of this theory to someone else as descriptively as possible. In doing so, we have to consider what is perceived as descriptive. Something is descriptive if one sees: it is similar to something else that I already know.

So we have to take the statements of the theory about the photons and find out in which respect and to which extent they behave like something which we already know.

One may object: But exactly that is the problem! There simply is nothing which behaves so strangely as the photons do! For the photons there does not exist any "it is similar to …" (apart from the souls). I think when we argue like that, we capitulate a little too quickly.

By the way, one has the impression not to have the same problems with the electron, and certainly not with the proton. But this is simply because the range of phenomena in which these particles behave like small individuals with a well-defined trajectory and shape is somewhat larger.

Origin:

The origin of the popular statement, photons are energy portions: a wrong understanding of the relation between physical quantity and physical system.

The origin of the reluctance with a statement about the shape of photons could be that the subject is dominated by the theorists, among whom the need for descriptiveness, is not as pronounced as for us school teachers.

Disposal:

We propose, at least for the teaching at school, to give the photon a little more vividness. We identify the size and the shape of the photon with its coherence region. It has thus an extension in the longitudinal and in the transverse direction. And this is different depending on the state of the light, thus depending on how one has produced it and what one has done to it afterwards. If the coherence length (i.e. the "temporal coherence") is very large, the momentum of the photons has a sharp value, the photons have a sharp wavelength. Large coherence width means that the transverse components of the momentum are small; one has broad wavefronts, which is important for interference experiments at the double slit.

Friedrich Herrmann

9

Solid State Physics

9.1 The semiconductor diode as a rectifier

Subject:

Text books often claim that the rectifier effect of a *pn* junction is due to the depletion layer at both sides of the contact surface between the *p* and the *n* region.

"If the *n* layer is connected with the plus terminal and the *p* layer with the minus terminal of the source, the depletion layer gets wider. The diode is blocking. If instead the *p* layer of the diode is connected with the plus terminal and the *n* layer with the minus terminal, free electrons and holes enter the depletion layer. Thereby this layer loses its effect and the diode becomes a conductor."

Deficiencies:

It is true that the thickness of the depletion zone changes as a function of the applied voltage. Thus, the above conclusion seems plausible. However, to infer the resistance from the density of the charge carriers is only correct if the charge carriers maintain their identity within the considered section of the diode. Such a conclusion is not valid if the charge carriers are subject to reactions. This is indeed the case for the *pn* junction. For forward polarity electrons and holes react to photons and phonons. For reverse polarity the reaction proceeds in the opposite direction, though with a much lower reaction rate, since at normal temperatures only few photons and phonons are present. It is this asymmetry of the reaction rate which is responsible for the asymmetry of the resistance.

The region in which the rectifying effect takes place is given by the diffusion length which, by the way, is 1000 times the thickness of the depletion layer.

Origin:

The traditional repugnance of the physicist against chemical arguments. It leads to the futile attempt to explain the processes in a semiconductor diode only with Ohm's Law and the laws of electrostatics, i.e. with the tools of electricity. Actually it is impossible to explain the diode, as well as the *pnp* and the *npn* transistor without recourse to the laws of chemistry. The explanation is most elegant when using the chemical potential gradient as a driving force which is analogue to the electric potential gradient.

Disposal:

The semiconductor diode working as a rectifier or a LED may be explained as follows: In forward polarization electrons from the *n* layer and holes from the *p* layer move towards the *pn* junction. There they react to create photons and phonons. The diode behaves like a closed switch. As an LED the diode is optimized in such a way that as few as possible phonons and as many as possible photons are created. In reversed polarization charge carriers should flow from the middle, i.e. the contact region outwards. Since there is no source of charge carriers at the *pn* contact no charge carriers can flow away. There is no electric current and no light is emitted. Only when examining more carefully one can notice that electrons and holes are indeed produced at a very low rate by the ambient thermal radiation. These charge carriers are responsible for the reverse current.

Friedrich Herrmann

9.2 The semiconductor diode as a solar cell

Subject:

In school books as well as in University text books one can find the assertion that in a solar cell the electric potential gradient in the space charge layer of the *pn* junction is the cause of the electric current generated by the cell:

"The separation of electrons and holes caused by the internal electric field within the depletion layer represents the generator effect."

"Due to electric forces the liberated electrons are pushed into the *n* layer and the holes into the *p* layer."

Deficiencies:

On a cursory inspection the statement seems plausible. The electric current caused by the solar cell needs a cause or a kind of driving force. Physicists know that an electric field represents such a cause. There is indeed an electric field within the diode, and its direction is the one that we need. Therefore, the physicist concludes, this field or the corresponding potential gradient is responsible for the electric current. Unfortunately the physicist has overlooked another fact. Never an electric potential gradient can be the cause of a stationary electric current. If we follow a (positive) charge carrier on its trajectory in the circuit, we observe that it goes just as much uphill (the potential hill) as downhill. Since in the load resistance it goes downhill, in the energy source it must necessarily go uphill. One can precisely recognize the energy source by the fact that the electric potential is higher at the terminal where the (positive) charge comes out than at the terminal where it enters the source.

The fact that on some sections within the energy source the charge goes downhill does not rebut the argument. In an electric circuit the electric potential changes each time that the material of the conductor changes. It does so in any conductor, even in a circuit without a battery. These potential steps add up to zero when going once around the whole circuit. That is why there is no need for care about this phenomenon.

Origin:

Again the desperate attempt to explain the working principle with the familiar tools of electricity, although with precisely these tools it can be seen that the argument is not correct.

Disposal:

The cause or driving force for a current of electric charge carriers can but must not be an electric potential gradient. Actually the cause of the electric current in the solar cell is a gradient of the chemical potential. Thus the solar cell is a close relative of the electrochemical cell.

Friedrich Herrmann

9.3 Field and diffusion current

Subject:

The electric potential gradient in a zero-current pn junction is the cause of a “field current”. The field current is compensated by the “diffusion current”. The diffusion current flows in the opposite direction and is a consequence of the concentration gradient of the charge carriers.

Deficiencies:

When in a conducting material an electric potential gradient exists and the chemical potential has the same value everywhere, there is a current of charge carriers. The charge carriers are “driven” by the electric potential gradient. An electric current will also flow when there is a chemical potential gradient (caused for instance by a concentration gradient) and the electric potential has the same value everywhere. In this case the “driving force” of the charge carriers is the chemical potential gradient. Thus, there are two possibilities to “pull at the particles”: the electric potential gradient pulls at the quantity Q, i.e. the electric charge, whereas the chemical potential gradient pulls at the quantity n , i.e. the amount of substance.

In general, both gradients are different from zero and the resulting driving force is due to both gradients. It can be described by means of the electrochemical potential η. The electrochemical potential is essentially the sum of the electric potential ϕ and the chemical potential μ:

$$\eta = \mu + z \cdot F \cdot \phi \, .$$

Then for the electric current density we get:

$$\vec{j} = -\frac{\sigma}{zF} \mathrm{grad} \eta \, .$$

σ is the electric conductivity, z the number of elementary charges of each charge carrier and F the Faraday constant.

It is possible that the gradient of the electrochemical potential is zero. That means that both driving forces are equal and opposite and thus compensate each another. In this case there is no electric current. We have “electrochemical equilibrium”.

Now, instead of saying, that a particle current can be driven in two ways, or that there are two “driving forces”, it is often said that the electric potential gradient causes a “field current” and the chemical potential gradient causes a “diffusion current”, and that both of these currents superpose to the total current. Then, in the case of the electrochemical equilibrium we would have two currents of the same magnitude flowing in opposite directions.

The problem with this interpretation is that each of these currents separately should produce entropy (and thereby heat). But we know that the total current is dissipationfree. There is no entropy production. And how are we supposed to imagine this situation on the microscopic scale: Should we believe that some of the charge carriers comply with the electric potential gradient and others with the concentration gradient? If we consider an arbitrarily chosen charge carrier: to which current does it belong?

That this description is inconvenient can also be seen by comparing the situation with a similar one, in which nobody would make a decomposition in opposite currents. Consider the air of the atmosphere. It is also subject to two driving forces: The gradient of the gravitational potential pulls the air molecules downward, the pressure gradient pulls them upward. When the air is motionless and the temperature uniform both driving forces are equal and opposite, they compensate each another. Why do we not say in this case that there is a field current downward and a diffusion current upward?

Origin:

Probably several causes add up: 1. The simple and powerful tool “chemical potential”, although introduced in physics more than a hundred years ago, is today nearly unknown and scarcely used. 2. The electrochemical potential is not taken seriously as a physical quantity.

Disposal:

There are two driving forces for charge carriers: an electric driving force that pulls at the electric charge and a chemical driving force that pulls at the amount of substance. Since the electric charge and the amount of substance are tightly coupled both potentials can be combined to one single potential, the electrochemical potential. The gradient of the electrochemical potential is responsible for the particle current.

Friedrich Herrmann and Peter Würfel

9.4 The photoelectric effect

Subject:

At school and university the photoelectric effect is demonstrated in order to prove the quantum nature of light. It allows for a simple measurement of Planck's constant with fairly good accuracy.

Fig. 1 shows the experiment schematically: Light is incident on a cathode that is made of a material with a low work function, typically an alkali metal.

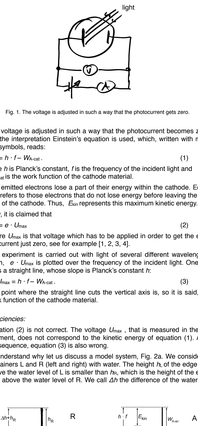


Fig. 1. The voltage is adjusted in such a way that the photocurrent gets zero.

The voltage is adjusted in such a way that the photocurrent becomes zero. For the interpretation Einstein's equation is used, which, written with modern symbols, reads:

$$E_{kin} = h \cdot f - W_{K\text{-}cat}. \tag{1}$$

Here *h* is Planck's constant, *f* is the frequency of the incident light and $W_{K\text{-}cat}$ is the work function of the cathode material.

The emitted electrons lose a part of their energy within the cathode. Equation refers to those electrons that do not lose energy before leaving the surface of the cathode. Thus, E_{kin} represents this maximum kinetic energy.

Now, it is claimed that

$$E_{kin} = e \cdot U_{max} \tag{2}$$

where U_{max} is that voltage which has to be applied in order to get the electric current just zero, see for example [1, 2, 3, 4].

The experiment is carried out with light of several different wavelengths. Then, $e \cdot U_{max}$ is plotted over the frequency of the incident light. One obtains a straight line, whose slope is Planck's constant *h*:

$$e \cdot U_{max} = h \cdot f - W_{K\text{-}cat}. \tag{3}$$

The point where the straight line cuts the vertical axis is, so it is said, the work function of the cathode material.

Deficiencies:

Equation (2) is not correct. The voltage U_{max}, that is measured in the experiment, does not correspond to the kinetic energy of equation (1). As a consequence, equation (3) is also wrong.

To understand why let us discuss a model system, Fig. 2a. We consider to containers L and R (left and right) with water. The height h_L of the edge of L above the water level of L is smaller than h_C, which is the height of the edge of R above the water level of R. We call Δh the difference of the water levels.

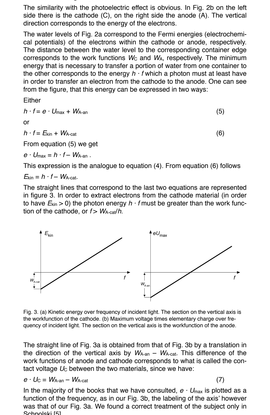


Fig. 2. (a) A portion of water of mass *m* is to be transferred from the container L at the left to container R at the right. For this process the energy $m \cdot g \cdot (\Delta h + h_R)$ is needed. (b) An electron is to be transferred from the cathode C to the anode A. For this process the energy $e U_{max}$ + $W_{K\text{-}cat}$ is needed.

Now we want to transfer a water portion of mass *m* from L to R. The energy that is necessary for this operation is determined by the difference of the water level in L and the height of the edge of R:

$$\text{Minimum transfer energy} = m \cdot g \cdot (\Delta h + h_R). \tag{4}$$

It is seen that height h_L does not enter into the result.

The similarity with the photoelectric effect is obvious. In Fig. 2b on the left side there is the cathode (C), on the right side the anode (A). The vertical direction corresponds to the energy of the electrons.

The water levels of Fig. 2a correspond to the Fermi energies (electrochemical potentials) of the electrons within the cathode or anode, respectively. The distance between the water level to the corresponding container edge corresponds to the work functions W_K and W_A, respectively. The minimum energy that is necessary to transfer a portion of water from one container to the other corresponds to the energy $h \cdot f$ which a photon must at least have in order to transfer an electron from the cathode to the anode. One sees from the figure, that this energy can be expressed in two ways:

Either

$$h \cdot f = e \cdot U_{max} + W_{K\text{-}an} \tag{5}$$

or

$$h \cdot f = E_{kin} + W_{K\text{-}cat} \tag{6}$$

From equation (5) we get

$$e \cdot U_{max} = h \cdot f - W_{K\text{-}an}.$$

This expression is the analogue to equation (4). From equation (6) follows $E_{kin} = h \cdot f - W_{K\text{-}cat}$.

The straight lines that correspond to the last two equations are represented in figure 3. In order to extract electrons from the cathode material (in order to have $E_{kin} > 0$) the photon energy $h \cdot f$ must be greater than the work function of the cathode, or $f > W_{K\text{-}cat}/h$.

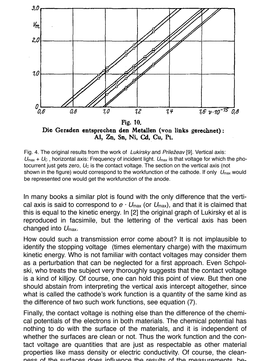


Fig. 3. (a) Kinetic energy over frequency of incident light. The section on the vertical axis is the work function of the cathode. (b) Maximum voltage times elementary charge over frequency of incident light. The section on the vertical axis is the workfunction of the anode.

The straight line of Fig. 3a is obtained from that of Fig. 3b by a translation in the direction of the vertical axis by $W_{K\text{-}an} - W_{K\text{-}cat}$. This difference of the work functions of anode and cathode corresponds to what is called the contact voltage U_C between the two materials, since we have:

$$e \cdot U_C = W_{K\text{-}an} - W_{K\text{-}cat} \tag{7}$$

In the majority of the books that we have consulted, $e \cdot U_{max}$ is plotted as a function of the frequency, as in our Fig. 3b, the labeling of the axis however was that of our Fig. 3a. We found a correct treatment of the subject only in Schpolski [5].

Even though one may follow our arguments, the following objections might arise: The experiment as it is carried out at the school or at the University lab, gives as a result the work function of the material of the cathode and not that of the anode. The latter would be much greater than the approximately 2 eV which are actually measured. The explanation for this strange behavior is that a small amount of Cesium (we suppose to have a Cesium cathode) has reached the surface of the anode. Actually the manufacturers of photocells advert to this effect. A sporadic covering of the anode's surface with Cesium is sufficient to allow all of the photoelectrons to enter into the anode material. Each spot of a material with a lower work function represents a potential minimum for the electrons so that the electrons voluntarily choose these locations to enter the anode material. According to the manufacturer's advice some photocells must be heated from time to time in order to clean the anode from the cathode material. Otherwise, the anode itself may begin to act as a source of photo electrons due to stray light.

Finally one might ask why the manufacturers make the cathode of a material with a small work function like Cesium, and why they do not use such a material for the anode. To answer this question we must remember what the photocells are produced for. Usually they are not made to enable physics teachers to measure Planck's constant. They are made to measure light intensities and for that purpose the applied voltage is in the other direction: not to stop the electrons but to extract them from the cathode. In order to be sensitive for light with long wavelengths the work function of the cathode must be small.

Origin:

Einstein's work on the effect is not an experimental work.

For a rather long time after his publication no experimental data were available. Einstein's was only interested in the explanation of the observation that the kinetic energy of the single electrons is independent of the light intensity, and that the number of the emitted electrons is proportional to the light intensity [6].

The effect was measured very thoroughly in the decades following Einstein's publication by various researchers. The most important work was done by Millikan [7, 8] and by Lukirsky and Prilezaev [9]. Figure 4 is from the publication of Lukirsky et al. It shows the kinetic energy E_{kin} of the emitted electrons as a function of the frequency of the incident light. According to equation (1) the axis intercept (not shown in the figure) on the vertical axis is to within a factor *e* equal to the work function of the cathode. The authors obtained the kinetic energy by adding the contact voltage between cathode and anode to the measured voltage U_{max}. They (just as Millikan) had measured the contact voltage independently.

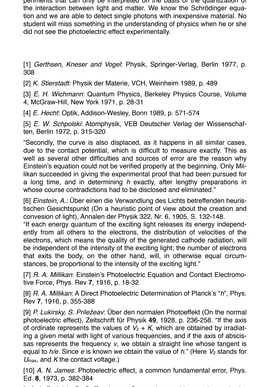


Fig. 4. The original results from the work of *Lukirsky* and *Prilezhen* [9]. Vertical axis: $eU_{max} + U_C$, horizontal axis: Frequency of incident light. U_{max} is that voltage for which the photocurrent just gets zero, U_C is the contact voltage. The section on the vertical axis (not shown in the figure) would correspond to the workfunction of the cathode. If only U_{max} would be represented one would get the workfunction of the anode.

In many books a similar plot is found with the only difference that the vertical axis is said to correspond to $e \cdot U_{max}$ (U_{max}), and that it is claimed that this is equal to the kinetic energy. In [2] the original graph of Lukirsky et al is reproduced in facsimile, but the lettering of the vertical axis has been changed into U_{max}.

How could such a transmission error come about? It is not implausible to identify the stopping voltage (times elementary charge) with the maximum kinetic energy. Who is not familiar with contact voltages may consider them as a perturbation that can be neglected for a first approach. Even Schpolski, who treats the subject very thoroughly suggests that the contact voltage is a kind of killjoy. Of course, one can hold this point of view. But then one should abstain from interpreting the vertical axis intercept altogether, since what is called the cathode's work function is a quantity of the same kind as the difference of two such work functions, see equation (7).

Finally, the contact voltage is nothing else than the difference of the chemical potentials of the electrons in both materials. The chemical potential has nothing to do with the surface of the materials, and it is independent of whether the surfaces are clean or not. Thus the work function and the contact voltage are quantities that are just as respectable as other material properties like mass density or electric conductivity. Of course, the cleanness of the surfaces does influence the results of the measurements, because if the surface is covered with dirt, one has to do with the chemical potential of the dirt instead of that of the bulk material.

Not only the origin of the error is interesting, but also the history of the vain efforts to correct it. In 1973 an article with the unambiguous title "Photoelectric effect, a common fundamental error" appeared in the English review *Physics Education* [10]. Three years later an article with the title "Concerning a widespread error in the description of the photoelectric effect" was published in the *American Journal of Physics* [11]. Its Authors seemed to ignore the British publication. In 1980 a similar article appeared in a German school science review with the featureless title "Work function and photoelectric effect" [12]. The author cites the *American* publication.

This story shows that an error can survive, even when a correction or revision is reminded. If the wrong idea is plausible and it is impossible does not cause too much harm, it seems the a correction is impossible.

Disposal:

Three possibilities.

1. Explain the effect correctly, for instance with the water model shown above.

2. Abstain from interpreting the axis intercept.

3. Abstain completely from carrying out and interpreting the experiment. For a scientist in the year 1910 or 1920 the experiment was important, it was a key experiment. Fortunately the students today must not acquire their knowledge under the same difficult conditions as students at this ancient time. We now know how the story ends and we know an infinity of other experiments that can only be interpreted on the basis of the quantization of the interaction between light and matter. We know the Schrödinger equation and we are able to detect single photons with inexpensive material. No student will miss something in the understanding of physics when he or she did not see the photoelectric effect experimentally.

[1] *Gerthsen, Kneser and Vogel*: Physik, Springer-Verlag, Berlin 1977, p. 308

[2] *K. Stierstadt*: Physik der Materie, VCH, Weinheim 1989, p. 489

[3] *E. H. Wichmann*: Quantum Physics, Berkeley Physics Course, Volume 4, McGraw-Hill, New York 1971, p. 28-31

[4] *E. Hecht*: Optik, Addison-Wesley, Bonn 1989, p. 571-574

[5] *E. W. Schpolski*: Atomphysik, VEB Deutscher Verlag der Wissenschaften, Berlin 1972, p. 315-320

"Secondly, the curve is also displaced, as it happens in all similar cases, due to the contact potential, which is difficult to measure exactly. This as well as several other difficulties and sources of error are the reason why Einstein's equation could not be verified properly at the beginning. Only Millikan succeeded in giving the experimental proof that had been pursued for a long time, and in determining *h* exactly, after lengthy preparations in whose course contradictions had to be disclosed and eliminated."

[6] *Einstein, A.*: Über einen die Verwandlung des Lichts betreffenden heuristischen Gesichtspunkt (On a heuristic point of view about the creation and conversion of light), Annalen der Physik 322, Nr. 6, 1905, S. 132-148

"If each energy quantum of the exciting light releases its energy independently from all others to the electrons, the distribution of velocities of the electrons, which means the quality of the generated cathode radiation, will be independent of the intensity of the exciting light; the number of electrons that exits the body, on the other hand, will, in otherwise equal circumstances, be proportional to the intensity of the exciting light."

[7] *R. A. Millikan*: Einstein's Photoelectric Equation and Contact Electromotive Force, Phys. Rev **7**, 1916, p. 18-32

[8] *R. A. Millikan*: A Direct Photoelectric Determination of Planck's "*h*", Phys. Rev **7**, 1916, p. 355-388

[9] *P. Lukirsky, S. Prilezhaev*: Über den normalen Photoeffekt (On the normal photoelectric effect), Zeitschrift für Physik **49**, 1928, p. 236-258. "If the axis of ordinate represents the values of $W_1 + K$, which are obtained by irradiating a given metal with light of various frequencies, and if the axis of abscissas represents the frequency *v*, we obtain a straight line whose tangent is equal to *h/e*. Since *e* is known we obtain the value of *h*." (Here V_0 stands for U_{max}, and *K* the contact voltage.)

[10] *A. N. James*: Photoelectric effect, a common fundamental error, Phys. Ed. **8**, 1973, p. 382-384

[11] *J. Rudnick, D. S. Tannhauser*. Concerning a widespread error in the description of the photoelectric effect, Am. J. Phys. **44**, 1976, p. 796-798

[12] *J. Strnad*: Die Austrittsarbeit beim Photoeffekt, Praxis der Naturwissenschaften – Physik, 1980, p. 343-344

Ralph von Baltz, Friedrich Herrmann and Michael Pohlig

9.5 Measuring Planck's constant by means of LED's

Subject:

Planck's constant can be measured by using light emitting diodes. The voltage applied to the LED is increased until the diode begins to emit light. The corresponding threshold voltage U_0 multiplied with the elementary charge is, so it is said, equal to the band gap energy and thus equal to the energy of the emitted photons. The experiment is carried out with various LED's which emit light with different frequencies.

Deficiencies:

There is no threshold voltage for the light that is emitted by the diode. The light intensity is proportional to the electric current in the diode. The electric current I as a function of the applied voltage U is in good approximation given by:

$$I = I_S \cdot \exp\left(\frac{eU}{\eta kT}\right) = I_S \cdot \exp\left(\frac{U}{U_T}\right). \qquad (1)$$

Here, k is the Boltzmann constant, T the absolute temperature and e the elementary charge. η is called the non-ideality factor whose value is between one and two. It would be equal to one if all the electron-hole pairs would recombine radiatively. η has no significance for the following considerations as long as it has the same value for all the diodes that are compared. I_S is the saturation current. It depends on the temperature and on the band gap energy E_g . The following proportionality holds:

$$I_S \sim A \cdot \exp\left(-\frac{E_g}{\eta kT}\right),$$

where A is the pn contact surface area. Apart from

$$U_T = \frac{\eta kT}{e}$$

there is no characteristic voltage in equation (1). However, U_T has nothing to do with the band gap [1, 2].

There is not a minimum voltage for which the diode begins to emit, since it emits always – but with an intensity that depends on the applied voltage. It even emits when U = 0 V, namely the thermal radiation. When the voltage increases, the intensity of the emitted light increases exponentially, whereas its spectral distribution does not change. It may surprise that the diode emits photons whose energy is roughly equal to the band gap energy, even though the energy eU supplied to the electron-hole pairs is smaller. Actually the diode would cool down a little when working with small applied voltages. It works as a Peltier element. Since this effect is small, it is covered by the unavoidable dissipative heat.

The procedure that is applied to get a "threshold voltage" U_0 is based on an illusion. Fig. 1 shows three times the same exponential characteristic, the difference between the three representations consisting only in the choice of the axis of ordinates. Each time that the scale of the vertical current axis is changed by a factor of 100, the graph is displaced horizontally by

$$\frac{kT}{e} \cdot \ln 10^2 = 0{,}119 \text{ Volt}$$

(We have admitted that η = 1.)

The "threshold voltage" which one would read from the graph changes by the same amount.

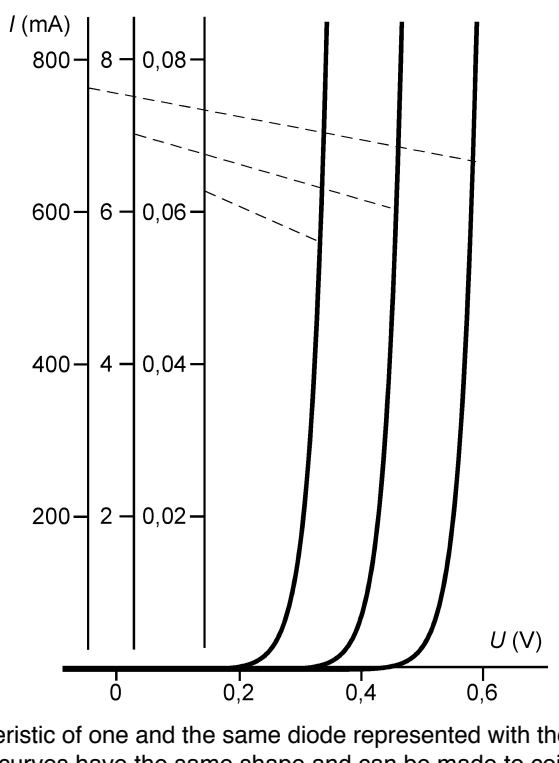


Fig. 1. Characteristic of one and the same diode represented with the current axis scaled differently. The curves have the same shape and can be made to coincide by a horizontal displacement.

Origin:

The experiment was introduced as a simple and inexpensive experiment for the physics lab at the high school and the university. The incorrect interpretation has a certain plausibility. Apparently, it was overlooked that a threshold voltage cannot be defined for an exponential function in principle.

Disposal:

Planck's constant can be determined by means of several diodes with different band gaps, Fig. 2.

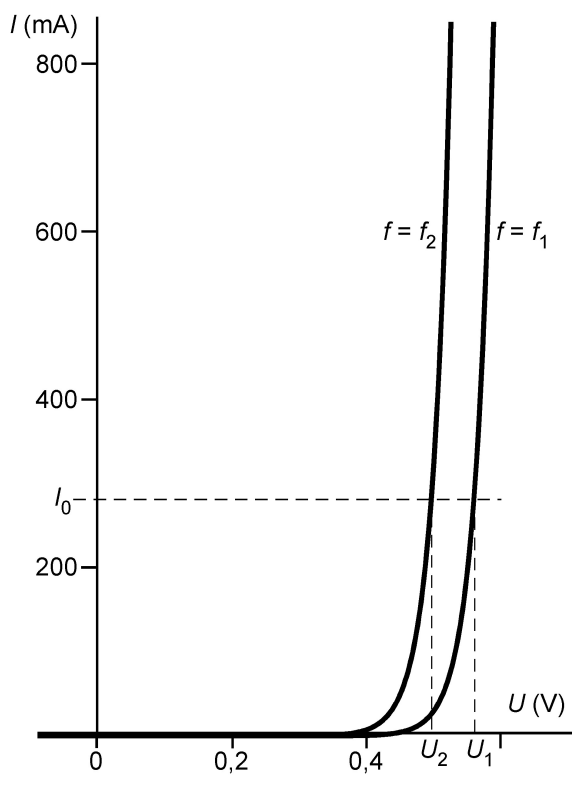


Fig. 2. Characteristics of two diodes, that emit light of different frequencies. The curves can be made to coincide by a horizontal displacement.

But there is a condition: the pn contact surface area must be the same for all of the diodes. If this is the case the corresponding characteristics are distinguished only in the factor [3]

$$\exp\left(-\frac{E_g}{\eta kT}\right).$$

The band gap energy E_g is related to the average frequency of the emitted light by

$E_g = h \cdot f$.

Thus, the horizontal distance between the two curves 1 and 2 is

$(E_{g1} - E_{g2})/e$.

We now choose an arbitrary value I_0 of the current and read the corresponding voltages U_i . We get

$U_1 - U_2 = (E_{g1} - E_{g2})/e$

or

$e(U_1 - U_2) = E_{g1} - E_{g2} = h(f_1 - f_2)$,

and thus

$$h = \frac{e(U_1 - U_2)}{f_1 - f_2}.$$

Notice that neither

$eU_1 = hf_1$

nor

$eU_2 = hf_2$

is valid separately.

When plotting eU_i over the frequency of the emitted light, one gets a straight line whose slope is equal to Planck's constant, fig. 3.

(Sometimes tangents at the points of equal current are drawn and the voltage is read where they cut the axis of abscissas. Obviously the value is the same as when reading the voltage directly as in Fig. 2, but it may give the illusion that this section represent something like a threshold voltage.)

Whether the straight line in Fig. 3 runs through the origin or not, depends only on the arbitrary choice of the value of the current I_0.

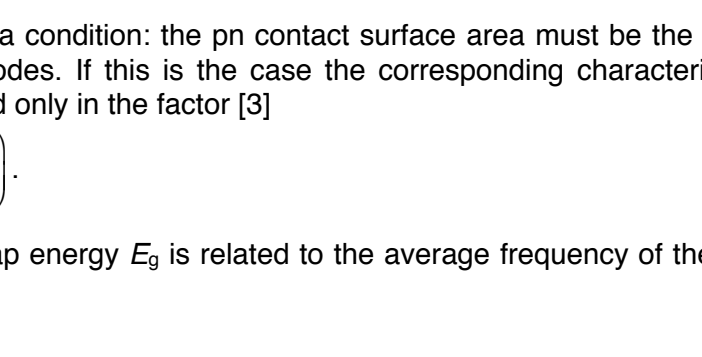


Fig. 3. For two (or more) diodes the voltage at I_0 multiplied by e is plotted over the frequency of the emitted light. The slope of the straight line is equal to the Planck constant.

Instead of reading the voltage values for a given current, one often uses another procedure: One chooses that voltage where the diode visibly begins to emit light. Since one automatically compares the light intensity with the ambient light, a voltage value can quite reliably be determined. By this procedure a current value is defined with sufficient precision that is the same for all diodes. This explains why the procedure gives satisfying results. However, the fact that the straight line obtained in this way often passes through the origin is pure coincidence.

[1] *Herrmann, F.* und *Schätzle, D.:* Question # 53. Measuring Planck's constant by means of an LED, Am. J. Phys. **64**, 1996, S. 1448

[2] *Morehouse, R.:* Answer to Question # 53. Measuring Planck's constant by means of an LED, Am. J. Phys. **66**, 1998, S. 12

[3] *Würfel, P.:* Physics of Solar Cells, Wiley-VCH, Weinheim 2009

Friedrich Herrmann and Peter Würfel

10

Nuclear Physics

10.1 Nuclear reactions and radioactivity

Subject:

The description of nuclear transformations, the discussion of measuring and detection processes for nuclear radiation.

Radioactive substances can emit three types of radiation: α, β and γ radiation. Nuclear transformation processes can be subdivided into radioactive decay, nuclear fission and nuclear reaction.

Deficiencies:

Nuclear physics is a real quarry of obsolete concepts. This becomes obvious when comparing the description of nuclear transformations with that of chemical reactions. And here we are already with the first deficiency. The similarity between normal chemistry and nuclear chemistry, or between the physics of the atomic shell and the physics of the nucleus goes much farther than it appears in many textbooks. By taking profit of this analogy nuclear physics could be conceptually simplified and by emphasizing the analogy learning could be facilitated.

In nuclear physics, concepts that existed already in chemistry, are sometimes introduced with a new name: What in chemistry is a monomolecular reaction is called in nuclear physics a decay o a spontaneous fission. The autocatalytic reaction of chemistry is a chain reaction in nuclear physics. The reaction rate is measured in chemistry in mol/s. In nuclear physics it has another name, namely activity, and is measured in Becquerel.

One should expect that the relation between the two measures is

$1\ \text{mol/s} = 6{,}02 \cdot 10^{23}\ \text{Bq}$.

However, practice is different:

$1\ \text{mol/s} = 6{,}02 \cdot 10^{23}\ \text{Bq} \cdot \text{mol}$.

Reaction equations are written differently in chemistry and nuclear physics. For example the reaction of the nuclides ${}^{9}_{4}\text{Be}$ and ${}^{4}_{2}\text{He}$ into ${}^{12}_{6}\text{C}$ and ${}^{1}_{0}\text{n}$ is written in nuclear physics as

${}^{9}_{4}\text{Be}\ (\alpha, n)\ {}^{12}_{6}\text{C}$.

whereas in chemistry the notation would be

${}^{9}_{4}\text{Be} + {}^{4}_{2}\text{He} \rightarrow {}^{12}_{6}\text{C} + {}^{1}_{0}\text{n}$.

In addition the notation of nuclear physics has a further inconvenient: It emphasizes an asymmetry between the reactants ${}^{9}_{4}\text{Be}$ and ${}^{4}_{2}\text{He}$, as well as between the products ${}^{12}_{6}\text{C}$ and ${}^{1}_{0}\text{n}$ which is not essential: the difference of the masses of ${}^{9}_{4}\text{Be}$ and ${}^{12}_{6}\text{C}$ on the one hand, and ${}^{4}_{2}\text{He}$ and ${}^{1}_{0}\text{n}$ on the other. Moreover, this notation is applicable only when there are exactly two reactants and two products.

Sometimes the same word is used in chemistry and in nuclear physics in different meanings. In nuclear physics in a reaction must participate at least two reactants, in chemistry not.

Who wants to learn nuclear physics has to do with particularly many technical terms. It is common to make unnecessary and unessential distinctions. An example: One insists to distinguish between natural and artificial radioactivity, i.e. between decay process of nuclides found in nature and man-made nuclides. Of course, also chemists could distinguish between natural and artificial compounds and their spontaneous decomposition. Fortunately they don‘t do so, because this distinction would not reflect anything essential.

It is also dispensable to give to some decay products “radiation names” in addition to their normal names. Moreover, the denominations α-, β- and γ-radiation suggest that between the corresponding particles there should be a similarity or an analogy, which is not the case. On the other hand, the relationship between a γ process and a photochemical reaction is usually not made evident.

Origin:

How did it come that the description of radiations dominates so strongly nuclear physics? Where does the proliferation of technical terms come from? Why do we spent so much teaching time for the description of radiation measuring processes?

The first, and for a long time the only known transformations of nuclei were related to “radiations”. Only thanks to the radiation it was possible to get information about a nuclear process, i.e. only by the fact that one of the reaction products had a low mass and thus takes over almost the whole energy released in the process. At the beginning, one observed the radiation and one did not yet know the nature of it. It was natural to give it a proper name. Moreover, at the times of the beginnings of nuclear physics radiation was in fashion. Several times the discovery of a new radiation was rewarded with a Nobel prize. Only slowly the similarities between processes of the atomic shell and the nucleus became apparent. Only decades later nuclear reactions with reaction rates as high as those known from chemistry have been observed or realized. Only in the 1920 it was understood that the sun is a nuclear reactor, and the first man-made reactor began to operate in 1942.

Disposal:

The disposal is not simple. It requires a comprehensive restructuration of the contents of nuclear physics. When doing so it is best to take chemistry as a model.

Friedrich Herrmann

10.2 Mass excess

Subject:

The mass of the atomic nucleus is smaller than the sum of the masses of its constituents. The difference is called the mass excess.

Deficiencies:

1. The term mass excess is introduced in the context of nuclear physics. However, the corresponding phenomenon also exists for the physics of the electronic shell. The mass of an atom is smaller than that of its constituent nuclei and electrons. Likewise, the mass of a molecule is smaller than the sum of the masses of the constituent atoms. The mass of two magnets that are attached to each other in such a way the opposite poles stick together is smaller than the mass of the magnets taken separately.

2. The designation suggests that there is a small deviation from the value which was to expect. However, when we go on to the nucleons and their constituent particles, the quarks, the mass excess is much greater than the mass of the constituents. Finally we have to be prepared to find out that the whole of the mass of any particle is “excess mass”.

3. The word “excess” normally expresses a nuisance. The mass excess can be a nuisance only for someone who does not know that there is a field which also has mass. Thus, the mass excess is not a deficiency. It rather puts in order a balance that otherwise would be incomplete.

Origin:

Once, the name was reasonable. Since Lavoisier discovered in 1772 the law of the conservation of mass, it was known that the mass of a substance is equal to the sum of the masses of its constituents. The law was proven to within the measuring accuracy and could be considered valid until about 1900.

The conservation of energy was discovered independently about a hundred years after the law of the conservation of mass. Only since 1905 we know that none of these two laws is valid in the original form, but that mass and energy are the same physical quantity and that the conservation principle holds only for this new energy-mass. The deviation between the mass of a nucleus and its constituent protons and neutrons would have seemed a miracle in prerelativistic times. After 1905 it was no more than a proof of Einstein’s theory. From a modern point of view, however, the name is misleading.

Disposal:

Treat the energy-mass equivalence as it corresponds to a modern point of view: as a matter of course. When doing so, there is not excess mass. No mass is exceeding.

Friedrich Herrmann

11

Chemistry

11.1 Physical and chemical processes

Subject:

The following citations are from chemistry text books:

“Processes, in which substances are transformed into other substances, are called chemical processes. In a physical process in general the state of a substance changes, whereas the essential properties remain unchanged: Sulfur remains sulfur, even if it is melted or vaporized.”

“Chemistry is the science of the substances and their changes. Physics on the contrary investigates the states and changes of the states of substances.”

Deficiencies:

1. The border line between physics and chemistry is drawn inappropriately: between the “chemical reaction” and the phase transition. However, these processes are tight relatives. It would be more convenient to stress their similarities. Both classes of processes can be described with the same methods and concepts. A chemical reaction can symbolically be written as

$A + B + C + \ldots \rightarrow R + S + T + ...$

A phase transition is that particular case, in which on the left and on the right side of the reaction arrow there is only one single substance, in symbols:

$A \rightarrow R$

This peculiarity, however, does not cause any essential difference in the mathematical treatment of the corresponding problem. The driving force for both types of processes is a difference of the chemical potentials between the reactants and the products. The value of the chemical potentials is taken from the same table in both cases. Also the heat balance is calculated by the same procedure and with values of the same table. In both cases there are exothermic and endothermic processes, both types of processes can be carried out reversibly and irreversibly.

2. If the definition is chosen as it is done here, one gets onto scrape anyway. The criterion for a process to be chemical is the formation of a new substance. But what is a new substance? Is a solution process physical or chemical? Is it chemical when hydration takes place and physical when not? Are gaseous, dissolved and crystalline NaCl different substances? And what about processes that occur in a solid material: The reaction of lattice vacancies with interstitial atoms or ions, the reaction of electrons with holes?

Origin:

It is not inequitable to explain what it is about when beginning a new subject area. It is conspicuous however, that chemistry text-books are particularly explicit in establishing a border toward physics. In physics text-books no corresponding efforts are made to demarcate the limit with chemistry. By the way, there are no tendencies to bulkhead between physics and electrical engineering, neither from one nor from the other side.

Disposal:

Instead of stressing the differences between phase transitions and “true” chemical reactions, it is better to treat these processes as particular cases of the same class of processes, to which many others also belong: the reaction of electrons with holes, the reaction of material substances with light, the reaction of interstitial atoms with lattice vacancies, the reaction of atomic nuclei ...

Friedrich Herrmann

11.2 Chemical equilibrium

Subject:

"In a chemical process, chemical equilibrium is the state in which the chemical activities or concentrations of the reactants and products have no net change over time. Usually, this state results when the forward chemical reactions proceed at the same rate as their reverse reactions. The rates of the forward and reverse reactions are generally not zero but, being equal, there are no net changes in any of the reactant or product concentrations. This process is known as dynamic equilibrium."

Deficiencies:

Consider two subsystems A and B. There are several types of equilibrium, namely as many as there are extensive variables X, which can be exchanged between A and B. (Only the energy does not define its own equilibrium, since it is exchanged together with any of the other extensive quantities.) To each of the extensive quantities X belongs an "energy-conjugated" intensive quantity ξ. If the systems A and B can exchange the extensive quantity X, this exchange comes to a halt only when the corresponding intensive variable has the same value for A and B, i.e. when $\xi_A = \xi_B$. Now the two subsystems are in equilibrium with regard to the exchange of X. The various equilibria are named according to the exchanged quantity. If two systems can exchange entropy, they are in "thermal equilibrium" when their temperatures are equal, i.e. when $T_A = T_B$. Two Systems that can exchange electric charge are in the state of "electric equilibrium" when their electric potentials are equal, i.e. when $\phi_A = \phi_B$. Two bodies which exchange momentum in a frictional process do so until their velocities have become equal, i.e. until there is "velocity equilibrium", or $v_A = v_B$. If in a chemical reaction the amounts of the substances A(1), A(2), A(3), ... can change at the expense of the amounts B(1), B(2), B(3), ..., the substances at the one side of the reaction equation are in "chemical equilibrium" with those of the other side, when the sum of the chemical potentials of the substances A(i) equals the sum of the potentials of substances B(k), i.e. when $\Sigma\mu_{A(i)} = \Sigma\mu_{B(k)}$.

When placing the chemical equilibrium in a broader framework, as we just have done, it is seen that it is not appropriate to emphasize that the chemical equilibrium is a *dynamic* equilibrium.

Consider again, for comparison, the electric equilibrium, and, to be concrete a piece of copper wire. Now imagine the wire consisting of two halves A and B. Sure enough they are in a state of electric equilibrium. It is common and reasonable to say that there is no electric current flowing between these two subsystems. If however, we describe this state in the same way as chemistry describes the chemical equilibrium, we would not be allowed to say that there is no current, since there is a continuous movement of electrons from A to B and from B to A, that results for a copper wire with a cross section of 1 mm^2 in an electric current of 10^8 A in one direction and a current of the same intensity in the other direction. Correspondingly, when no wind is blowing we would not be allowed to say that the air is at rest but we should say that we have a mass flow of about 100 $kg/(m^2 \cdot s)$ to the right and a similar flow to the left, and also currents of the same intensity back and forth and upwards and downwards. Similar conclusions would be drawn for thermal equilibrium, where we have currents of phonons in all directions, or for the velocity equilibrium related to continuous flows of momentum in opposite directions.

Of course, there is nothing incorrect in considering a phenomenon at the microscopic level. But, first, there is no essential difference in this respect between chemical equilibrium and other equilibria, for which nobody emphasizes that the equilibrium is a "dynamic equilibrium". And second, one is stressing something that easily leads to a misconception. If we say that in a copper wire in which no net current is flowing, "in reality" there are two counter-flowing currents, should the wire not heat up? Correspondingly, one could ask, why the two counter-running chemical reactions are not dissipative? Obviously, these problems are home-made. One is intermixing the arguments of two levels of description, the microscopic and the macroscopic.

Origin:

The description of chemical reactions on the simple, phenomenological level by means of the chemical potential has never won recognition. This is different from the other, physical phenomena mentioned above. There it is understood to describe a heat transport as caused by a temperature difference or an electric current by an electric potential gradient. The microscopic interpretation of these processes is done later in the context of atomic and solid state physics. Chemistry teaching begins at the molecular level, on which the simple and elegant thermodynamical quantities need a complicated interpretation.

Disposal:

Say that, when chemical equilibrium is reached, the reaction has come to a halt. This does not hinder us to consider the continuous forth and back reaction at a later advanced state. Just as we say that in the state of electric equilibrium we say that no electric current is flowing, and that this does not hinder us, to explain this state later on microscopically by the symmetry of the Fermi surface.

Friedrich Herrmann

11.3 Electrochemical cells

Subject:

Apart from electric generators, electrochemical cells are the most important electric energy sources. Historically they were the first technical electric energy sources. There is not doubt that they should be treated in the secondary physics education.

Deficiencies:

They are not found in the physics curriculum. Why? The opinion of the physics teachers may be: “There is not much to understand. All there is to do is to learn by heart the various reactions occurring at the electrodes. These are different according to the type of the cell. Thus, it is a subject typical for the chemistry class.”

The impression one gets when consulting the chemistry text book seems to confirm this conclusion. One is lavished with so many details and technical terms that at the end one is unable to notice that the question has remained unanswered – a procedure, that we scientists often reproach to the humanists. The quantity, that would allow for an explanation which is independent of the details and the peculiarities of a particular reaction, i.e. the chemical potential is even not introduced – neither in the physics nor in the chemistry lessons.

Moreover, the subject does not belong exclusively into chemistry. It belongs also to physics, because the electrochemical cell can be explained with methods, which are typical for physics and second, because in general the details of a particular reaction do not matter.

A comparison with the treatment of another class of electric energy sources is advisable. We treat the generator by showing the basic principle by means of a simple model experiment. In this way, the effect which is the base of all electric generators can be understood. The many variants of realistic technical generators are at best betoken. We should proceed in a similar way when treating electrochemical cells. The general working principle should be in the foreground.

Origin:

The fact that the chemical potential is not used.

Gibbs’s fundamental equation

$dE = TdS - pdV + \mathbf{v}d\mathbf{p} + \mu\, dn + \phi dQ - \ldots$

tells us which physical quantities are needed to describe energy exchanges: the thermodynamical quantities temperature T, entropy S, pressure p and volume V, the mechanical quantities velocity $\mathbf{v}$, and momentum $\mathbf{p}$, the chemical quantities chemical potential μ and amount of substance n, the electrical quantities electric potential ϕ and electric charge Q, etc. It happens that two of these quantities are almost not in use, just as if they were off-limits: entropy and chemical potential. For that a high price has to be paid: Either one helps himself with cumbersome surrogates – as for instance the enthalpy (instead of the entropy) as a measure for heat, that suits not really well, or the energy devaluation in order to describe entropy production –, or one simply eliminates those subjects from the curriculum, that could be explained by using these quantities, – as in the case of the electrochemical cell.

Disposal:

Who is not afraid of the chemical potential explains the electrochemical cell in the following way, Fig. 1:

The substances A and B can react to C:

$A + B \rightarrow C$

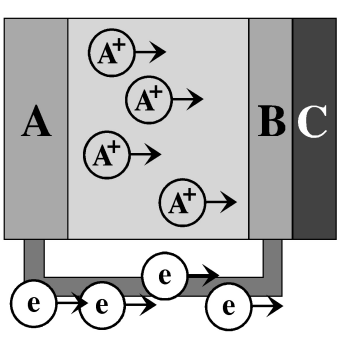


Fig. 1. The salt solution is a conductor only for A^+-ions, the copper wire only for electrons.

The reaction is driven by the chemical potential difference

$\Delta\mu = (\mu_A + \mu_B) - \mu_C$.

The chemical potentials of the various substances are tabulated. When the reaction extent is ξ, the energy

$E = \Delta\mu \cdot \xi$

is released as electric energy. How does the cell work?

As long as the reactants A and B are separated in space one from the other, they cannot react. The reaction has an infinite *reaction resistance*, just as no electric current flows between two bodies with different electric potentials, as long as there is no conductor between them. If A is a gas and we connect the containers with a pipe, A can flow to B and the reaction can begin. In this case, however, the whole energy, that is released will be used, or misused, for the production of heat. We now establish a connection of a particular kind, see the Figure.

A and B are not joined by one connection but by two. One of them – a salt solution (the so called electrolyte) – is permeable or is a conductor only for A^+ ions, but not for electrons e. The other one – a copper wire – is a conductor for electrons and a non-conductor for A^+ ions. Now, A can proceed to B only by separating into A^+ and e. A^+ goes through the A^+-conductor and e through the e-conductor. When arrived at B, they can react with B to C. For the moment, all the released energy would again only serve to produce heat. However, we now have the possibility to use one of the two currents to drive something. It is more comfortable to choose the electron current. In this way all of the released energy can be tapped.

The out-coming energy can be expressed by means of the voltage $\Delta\phi$ and the electric charge Q. We have:

$\Delta\phi \cdot Q = \Delta\mu \cdot \xi$.

Since $Q = z \cdot F \cdot \xi$ (z is a small integer, depending on the nature of the reaction, and F is the Faraday constant), we get the voltage of the cell:

$$\Delta\phi = \frac{\Delta\mu}{z \cdot F}.$$

Summarizing: The electric charge has to go up-hill within the cell, i.e. against its own natural tendency. In order to do so the charge carriers need another driving force. This is the chemical potential difference. Thus the charge carriers go within the cell up-hill the electric potential mountain and down-hill the chemical potential mountain.

Friedrich Herrmann

11.4 Electrolytes and doped semiconductors

Subject:

An electrolyte, we learn, is a material that is decomposed when an electric current is flowing through it.

“An electrolyte is a substance, that is at least partially ionized, thereby conducts the electric current, and decomposes.”

“Solutions, that conduct the electric current, like hydrochloric acid and thereby decompose are called electrolytes.”

“Liquids can be conducting or non-conducting, i.e. dissociating or non-dissociating. The conductors are called electrolytes. They are decomposed by the electric current.”

Deficiencies:

When a substance is decomposed electrolytically, –the process is called electrolysis– an electric current is flowing through the substance, which has to exist as a solution or a melt. The substance is called an electrolyte, and the designation is reasonable in this case (λύσις *lysis* solution).

The designation is not so reasonable when one is referring to the medium between the electrodes of a galvanic cell, since in this case the intention is not to decompose a substance. The role of the solution between the electrodes is rather that of a “selective conductor”: the “electrolyte” has to be a conductor for certain ions, and must be insulating for electrons. The unhappy designation is probably one of the reasons why hardly any student understands the working principle of the galvanic cell.

Here in brief how the galvanic cell works:

To create an *electric* potential difference one takes profit of the *chemical* potential difference of a reaction with two reactants. Let us write the reaction equation as

$$A + B \rightarrow AB\ .$$

The cell is constructed in such a way that the reaction cannot run at first because the two reactants are spatially separated. A can get to B, but only by splitting into two parts –in A-ions and electrons–

$$A \rightarrow A^+ + e^-\ ,$$

These two species go from A to B on separate ways. The A^+ ions go via the ion conductor (the “electrolyte“), whereas the electrons take the outer part of the circuit, in general a copper wire. The energy load is connected in the outer part of the circuit.

Regarding the electrolyte what matters is the fact that it is a conductor for A^+ ions and an insulator for electrons.

The working of solar cells is based on the same mechanism: the n-doped part of the semiconductor is a conductor for electrons and an insulator for holes, the p-doped part is a conductor for holes and an insulator for electrons. The free electrons and holes that are created by the light in a high concentration and thus with a high chemical potential have a tendency to go to places with a lower chemical potential, i.e. to leave the place where they are created. The electrons can get away only via the n-doped material, the holes can escape only through the p-doped material. Thus, again one takes profit of the selective conductivity of a material. Although the role of the n- and p-doped material is the same as that of the electrolyte in the galvanic cell, the word electrolyte is not employed.

Origin:

Just as other special disciplines electrochemistry has developed its own jargon, which for the concerned specialists may be quite useful, but which can represent an unnecessary obstacle for the beginner.

Disposal:

Introduce the substance between the two electrodes as a conductor for ions and an insulator for electrons, just as the copper wire of the external circuit has to be a conductor for electrons and an insulator for ions. Correspondingly one characterizes a n-doped semiconductor as a material that is a conductor for electrons and an insulator for holes, and analogically the p-doped material. In this context electrolysis should not be mentioned, and the word electrolyte not be used.

Friedrich Herrmann

11.5 The drive of substance flows – particle number density or chemical potential?

Subject:

In physics and chemistry, we are confronted with “currents” or “flows” of physical quantities of all kinds: electric currents (= currents of the electric charge), mass and volume currents, and also substance currents, or better: currents of the amount of substance, because the flowing quantity here is the amount of substance. Each current can be hindered by a “resistance”. It is then said to be dissipative. In this case it needs a “drive” or “driving force”: in the electrical case it is an electric potential gradient, the mass flow needs a gravitational potential gradient, a heat flow needs a temperature gradient. For the flow of the amount of substance, the gradient of the particle number density is usually introduced as the drive quantity. The transport itself is then called diffusion. It is said that a substance diffuses from places with a higher particle density to places with a lower particle density.

Deficiencies:

First of all, a detail: The physical quantity, the density of which is at issue here, is the amount of substance. It is a basic quantity of the SI unit system. If one uses the particle number density instead, it is like using the elementary charge number density instead of the electric charge density. Just as it can sometimes be interesting to look at the swarming electrons, in the case of diffusion it may sometimes be practical to look at the swarming of particles. For most practical questions, however, it is a good idea to operate with the charge density or with the density of the amount of substance, respectively. The sentence one would like to pronounce in connection with diffusion would then rather be: The substance diffuses from the higher to the lower density of the amount of substance.

Now our actual topic.

The quantitative formulation of the statement is Fick’s first law; in modern spelling:

$$\vec{j}_n = -D \cdot \operatorname{grad} \rho_n \tag{1}$$

ρ_n is the molar density (density of the amount of substance n) and $\vec{j}_n$ the flow density of the amount of substance. The factor D in front of the gradient is the diffusion constant. For ideal gases, it is independent of the molar density.

In this description of the diffusion, the gradient of the molar density appears as the cause or as the drive of the substance flow.

One can see that the equation belongs to a series of several other equations which all play an important role in the thermodynamics of irreversible processes. They describe flows or currents of extensive physical quantities where a resistance has to be overcome, so-called dissipative currents, i.e. currents with entropy production.

A well-known example is the expression for the electric current density $\vec{j}_Q$:

$$\vec{j}_Q = -\sigma \cdot \operatorname{grad} \varphi \tag{2}$$

Here φ is the electric potential, and σ the electric conductivity.

Equation (1) tells us that the substance current flows from high to low molar density, but equation (2) does not tell us that the electric current flows from high to low charge density. This may sometimes be the case, but only sometimes.

As far as the molar current is concerned, in certain cases the molar density can be considered as the driving force, namely whenever the system in which the current flows is homogeneous (apart from the inhomogeneity of the molar density) and when the diffusing substance follows the ideal gas equation. In general, however, the appropriate measure for the drive is the chemical potential μ, which is also formally analogous to the other cases.

Instead of equation (1) one has then:

$$\vec{j}_n = -K \cdot \operatorname{grad} \mu \tag{3}$$

In this formulation the equation applies always, i.e. not only for ideal gases and homogenous systems (provided of course that there is no other drive, so we are not dealing with coupled flows).

In the case of the ideal gas

$$\mu = \mu_0 + RT \ln \frac{\rho_n}{\rho_{n0}}$$

and the factor K in equation (3) is proportional to the mass density:

$$K = \frac{D\rho_n}{RT}$$

But if D is independent of the mass density, is not equation (1), at least for ideal gases, the simpler, the more beautiful equation? The simpler yes, the more beautiful not.

Because if one interprets the equation as it is reasonable, namely that the gradient represents the driving force for the current, then equation (1) makes a statement that does not fit into the picture: For a given drive one would expect that the current is proportional to the density of the “flowing quantity”. This is true in the electrical case (and also in the thermal case). The electric conductivity in equation (2) is known to be proportional to the charge density of the moving charge carriers.

Origin:

Fick’s first law, Equation (1), was published in 1855, i. e. before Gibbs (1873) introduced the chemical potential. One can see here, as well as in many other places of the of physics syllabus: Once introduced, nothing can be changed.

Disposal:

Introduce the chemical potential, an easy to understand, benign and universally usable quantity. Then Fick’s law can be written in the form of equation (3) and its resemblance with the corresponding electrical law becomes visible. By the way, it very nicely says in Wikipedia: “At a fixed pressure p and a fixed temperature T, the gradient of the chemical potential μ is the driving force of the substance flow from the point of view of thermodynamics”.

Friedrich Herrmann

11.6 The drive of substance flows – substance flows across phase boundaries

Subject:

Substance flows or transports are ubiquitous processes. A certain class of such transports is diffusion. It is rarely mentioned in physics schoolbooks – sometimes only in connection with the semiconductor diode. In chemistry books it is treated throughout. One learns, for example:

> "Diffusion is the spontaneous mixing of two substances. This mixing is due to the continuous random motion of the particles of the substances."

Or also:

> "A substance diffuses from places with a high particle density to places with a low particle density."

Deficiencies:

It is a pity that transports of substances, especially when teaching physics, are so casually treated. The fact that a substance (or its particles) spontaneously goes from high to low concentration only applies to the special case of homogeneous systems. In general, it can also be the other way round.

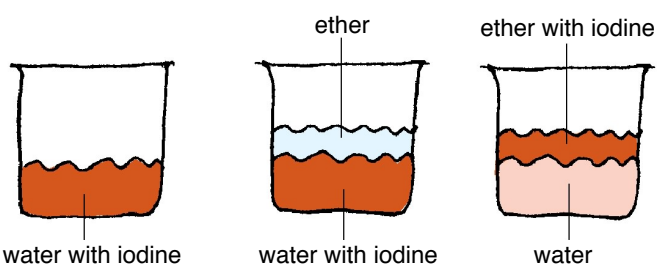


Fig. 1 The iodine goes from the water into the ether, following the chemical potential gradient, but against the concentration gradient.

Here (Fig. 1), a simple experiment showing a substance transport from the low to the high concentration ("particle number density"): Dissolve some iodine in water; the solution is brown. Overlay with some ether; the ether is colorless. If you stir the two liquids vigorously and wait a while, they separate again and form the same layering as before: above the ether, below the water – but with one difference: Now the ether is dark brown and the water only light brown. Most of the iodine is now dissolved in the ether. At first, following the concentration gradient, the iodine went from the water into the ether, but then it went on against the concentration gradient. At the end, the chemical potential of the iodine within the ether and within the water is equal, even though the concentration within the ether is higher.

One can see: Not the concentration gradient is responsible for a substance transport, but the gradient of the chemical potential μ. Concentration gradients and chemical potential gradients may correlate in special cases – but only in special cases.

Knowing this rule, which God knows is not complicated, opens up the possibility of explaining many phenomena that are otherwise only described.

- How can the fish breathe through their gills? How is it that oxygen is in the water? What is it doing there anyway? Isn't it gradually consumed by the fish? There is in good approximation a chemical equilibrium between the oxygen in the water and the oxygen in the air: the chemical potential of the oxygen in the water is equal to that in the air. When the fish consume some, new oxygen comes in from the air. So there is always oxygen in the water, the fish don't have to worry.
- Why does the oxygen in our lungs go from the air into the blood? Because its chemical potential in the air (–3.88 kJ/mol) is higher than in the blood entering the lungs through the pulmonary arteries (–7.30 kJ/mol). This increases the potential in the blood to –5.03 kJ/mol when it exits the pulmonary veins.
- Open a bottle that contains carbonated mineral water but is only about half full, blow away the CO_2 above the water surface, close the bottle again and shake. When you open it again, it hisses, because overpressure has built up since CO_2 has gone from the water into the gas area. The concentration in the gas phase is higher than in the liquid phase. What was the driving force behind this substance transport? Again the chemical potential difference.

It is the power of the chemical potential that it describes not only the diffusion within homogeneous phases, but also the substance transport under arbitrarily complicated boundary conditions. And not only that: The direction of a chemical reaction is also determined by the chemical potential, the same quantity that is responsible for diffusion, for each phase transition and for each substance transport across a phase boundary.

Origin:

1. The chemical potential is one of the simplest, most tangible quantities ever. Nevertheless, it had a difficult fate. In chemistry, because chemists fell in love with the non-intuitive thermodynamic potentials, i.e. the Legendre transforms of the function $U(n,V,S)$. While the energy as a function of the extensive quantities still represents a quite vivid function, the others, namely H, F, G and several more, are so abstract that one can only entrust oneself to them in blind flight: i.e. one leaves the solution of the problem to mathematics and hopes that the result is correct.

2. There is perhaps another reason why physicists do not like the quantity μ: it has the wrong name. Why should a physicist use a variable with a reference to chemistry in its name?

Disposal:

To put it somewhat casually: Just as one introduces the electric potential as energy per charge or the absolute temperature as energy per entropy, one can introduce the chemical potential as energy per amount of substance. It fits perfectly into the picture used elsewhere.

Friedrich Herrmann

12

Optics

12.1 Geometrical optics – wave optics

Subject:

"The limiting case of wave optics in which the wavelength $\lambda \to 0$ is called geometrical optics (or ray optics). In geometrical optics the wave nature of the light and the phenomenon of diffraction is not taken into account."

"If the wave length of the radiation energy decreases in comparison with the physical dimensions of the optical system, the effects of diffraction become less significant. In the limit of this concept, when $\lambda \to 0$, the straight propagation in homogeneous media is valid, and we obtain the idealized domain of geometrical optics."

„In cases where the wavelength is small compared to other length scales in a physical system, light waves can be modeled by light rays, moving on straight-line trajectories and representing the direction of a propagating light wave."

Deficiencies:

If the wave length is small, the condition for straight propagation of the light is fulfilled. However, we all know the experiment with the Fresnel's double mirror: An enlarged laser beam is sent on the double mirror. The two reflected partial beams generate interference fringes on a screen. Although the condition that the wavelength is small compared to "other length scales", a typical wave phenomenon is observed. To get rid of the wave properties of the light yet another condition has to be fulfilled: The light must be sufficiently temporally incoherent.

Origin:

Geometrical optics has developed rather independently from wave optics. The aim was the realization of optical instruments, that work with the light of the sun, of stars and of incandescent lamps. Due to the thermodynamical equilibrium of these light sources the emitted light has maximum entropy, and is therefore perfectly (temporally) incoherent. When two light beams cross each other or superpose, the average energy current densities can be added. The field strengths, that would have to be added in wave optics are not known anyway.

Disposal:

Name two properties that the light must have in order to behave according to the rules of geometrical optics: small wavelength and incoherence.

Friedrich Herrmann

12.2 The components of light

Subject:

One often says that white light consists of components with different wavelengths or frequencies:

> "White light consists of a mixture of all wavelengths of light."
>
> "Why does white light consist of many different wavelengths?"
>
> "Dispersion is the separation of white light into its constituent colors."
>
> "Light consists of photons; a photon has a well-defined wavelength; light consists of contributions of different wavelengths."

Deficiencies:

We prescind from solecisms like "white light consists of different *wavelengths*" or "white light consists of different *colors*", i.e. formulations that must hurt everybody who is accustomed to conceptual clearness. The subject had been discussed in our article Nr. 144.

Light can be described in various ways. In other words, there exist various theories of the light: geometrical optics, which deals with optical imaging, nonimaging optics which is mainly about energy flow distributions in light fields, wave optics, which is essentially an electromagnetic description, the thermodynamical description of light, which is important for instance for determining the efficiency of solar cells, and finally quantum optics. Non of these theories is incorrect – of course not. Nor can it be said that one of them is better than the other. Which theory one chooses only depends on the kind of problem one wants to solve.

In the following we limit ourselves on that descriptions which are important in school: wave optics and the thermodynamics of light.

Let us come back to our quotes, which can be considered as typical statements about the light: According to them light *consists* of components of various wavelengths.

However, light does not consist of sine components, if the wording "consist of" is understood in the meaning of the colloquial language. "Consist of" means: the components of an object are contained within the object and can be recognized in it. Instead of "consist of" one should better say: "can be decomposed in" (contributions of different wavelengths).

Actually, the light can be decomposed not only in sine contributions. There are many other possible decompositions.

Would we say, that the waves at the surface of the sea consist of waves of different wavelengths? If somebody would say so, probably we would spontaneously object: But you see that these are not sine waves – until we remember that the chaotic movement of the water of the sea can be Fourier decomposed just as the light of the sun.

There is reason to fear that in the mind of our pupils we create an idea about the light that corresponds to Fig. 1, where the wave trains may possibly be identified with the photons. (Perhaps one says that a photon has a well-defined wavelength, see our forth quote. In this case the photon should be of infinite length. At the same time one suggests that the photon is not very long, but rather point-like, and one says nothing about its width.)

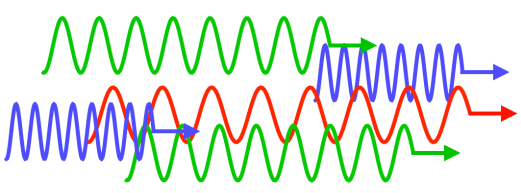


Fig. 1. Constituents of the light? Photons?

It would be more appropriate to begin the description of the state of the electromagnetic field that we call sunlight without reference to the Fourier decomposition: the field is in a state of maximum disorder, or maximum entropy. One can also say, in a state of maximum washing out in phase space with maximum incoherence. The time dependence of the amount of the field strength looks roughly like Fig. 2.

Fig. 2. White light: Field strength at a given point as a function of time

As a physicist one tends to consider this field as ugly. In communications engineering one would call such a time function noise, and noise is a phenomenon that is preferably to be avoided or eliminated. Isn't it true that a monochromatic plane wave is best for experimenting? Also the mathematical description of the monochromatic wave is much simpler than that of the chaotic white light, isn't it? Not necessarily. The chaotic or "thermal" light is the most simple according to another criterion. It can be described by means of only two variables: its temperature and its chemical potential. And in most of the situations one is interested in, the chemical potential is zero. Thus, what from one point of view appears as a maximum of complexity, is from another one particularly simple. Philosophers of science describe this simplicity that grows out of complexity as *emergence*.

Origin:

In the first place the prism: If behind it light of the various colors come out, the conclusion appears obvious, that these contributions had been the constituents of the light that entered the prism. A similar flippancy in the use of the "consists of" can be observed in other contexts, for example in atomic physics: It is sometimes said that the electronic shell of an atom *consists* of certain, well-defined orbitals. However, the shell can also be decomposed is a different way, and this is indeed done, when it is convenient. These parts or contributions is then given the somewhat daunting name hybrid orbitals. To the student it appears as something that is fundamentally new, and that is hard to understand, maybe even as a sleight of hand. Actually, the entire atomic-shell-cake has simply been divided in pieces of another shape.

It may be that yet another factor contributes to the idea, that white light consists of sine waves. The statement "Light is an electromagnetic wave" is not incorrect in the sense of physics. However, in our colloquial language by a wave we understand a periodic phenomenon. A function like that of Fig. 2 would not be called a wave.

Disposal:

Carefulness with our wording. Make clear that the spectral decomposition is only one among many others.

Do not introduce the sine waves as the real nature of light. Do also show images of non-sine light distributions.

And finally: A little thermodynamics is not harmful.

Friedrich Herrmann

[1] F. Herrmann: *Historical burdens on physics*, 43, The field of permanent magnets

[2] F. Herrmann: *Historical burdens on physics*, 60. Inductivity

[3] A. Sommerfeld: Elektrodynamik, 4. Auflage, Akademische Verlagsgesellschaft Geest & Portig, Leipzig, 1964, Vorwort, S. VI

12.3 Imaging and non-imaging optics

Subject:

When learning geometrical optics, its only goal seems to be to create optical images. It is used in the construction of optical instruments such as magnifying glasses, spectacles, microscopes and telescopes. In any case, as much light as possible that emanates from one point of the object should meet again in one image point.

Deficiencies:

In a successful optical imaging, the points of an object are “imaged” on points in the image plane: As much light as possible, which emanates from an object point, should merge into one image point. If possible, all light rays that pass through the optical system should converge again in the image point. It is also expected that the image is not distorted, i.e. that the ratios between the angles at which the image points appear from a point on the optical axis are equal to the ratios at which the object points would be seen.

If one regards the process of imaging as an energy transport, one can also say: One realizes an energy transport with light, which has to fulfill an additional condition.

If one looks around, where in nature and technology light transports are realized, then one notes that the transports, which are associated to an optical imaging, are only a special case. They are important in certain contexts – namely always when the transmission of information is desired –, but not in others.

If one restricts oneself to the requirement to transport energy with light from one place to another without creating an optical image, one discovers that the demands made on the optical system are not simply less restrictive, but completely different. We are dealing with the field of non-imaging optics.

Like imaging optics, non-imaging optics belongs to the field of geometric optics. Its aim is to bring as much light as possible from a source, usually a luminous surface, to a receiver. Non-imaging optics is responsible when the problem is illumination, concentration of light or the collection of light.

It is a pity that imaging optics has become so dominant in schools and universities that there is no time left for non-imaging optics. The impression results that problems related to lighting or the concentration of light are simply a somewhat coarse application of imaging optics. The best device that solves the problem, it may be thought, is a multi-lens optical system that corrects lens errors as far as possible. With this idea in mind, however, one would be far off the mark. The new question leads to a completely different optics, in which different laws and rules are relevant, and in which a well corrected lens system is a bad solution.

Origin:

The questions of non-imaging optics probably arose later than those of imaging optics. In addition, non-imaging optics, as well as numerous other technical applications of physics, were early spun off from physics, resulting in the special discipline of lighting technology, which at the university is possibly assigned to the electrotechnical faculty.

Disposal:

The problem to be investigated is not the same as with imaging optics. The problem is to transfer as much light as possible from one surface to another. A typical device of non-imaging optics is the light concentrator. In a concentrator, light enters through an opening with the surface area A_1 and exits through another opening with the surface area A_2, Fig. 1.

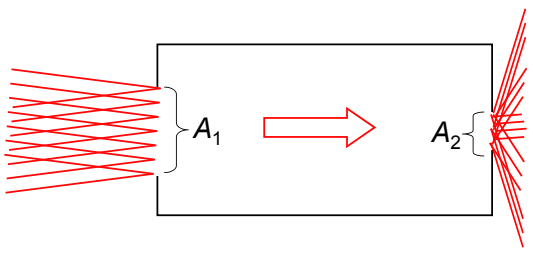


Fig. 1. Concentrator schematically

The most important law to consider when designing the concentrator is the sine condition, also called Abbe’s sine condition, which is nothing else than the 2nd law:

$$A \cdot \sin^2 \alpha = \text{const}$$

Here A is the cross-sectional area of the light beam and α is the aperture angle of the light distribution at each point of the cross-sectional area. The sine condition tells us that the “disorder in the position” (i.e. the area) can only be reduced if the “angular disorder” increases: Entropy cannot be destroyed.

Even before we start to ask how to construct a concentrator, we can make an important statement about the concentration factor

$$c = \frac{A_1}{A_2}$$

With the sine condition we can write:

$$c = \frac{A_1}{A_2} = \frac{\sin^2 \alpha_2}{\sin^2 \alpha_1}$$

Now the aperture angle α_2 at the output of the concentrator cannot be greater than 90°. The concentration factor can therefore at most assume the value.

$$c_{max} = \frac{1}{\sin^2 \alpha_1}$$

This short calculation already contains a lot of fundamental and at the same time plausible physics:

The smaller the aperture angle of the incoming light, the more it can be concentrated.

For diffuse light, i.e. light with $\alpha_1 = 90°$, the concentration factor becomes one; the light cannot be concentrated.

For the sun, $\alpha_1 = 0.266°$. This results in a maximum concentration factor of 46 400.

An optimally calculated concentrator achieves 96% of the theoretically maximum possible concentration. However, it is not worth building such a concentrator at all, because a simple cone-shaped mirrored funnel already achieves 92%.

One might expect that a corrected lens can do even more. In fact, a lens, whether corrected or not (aperture ratio 1.7), only achieves 10% of this.

The latter statement shows that non-imaging optics is not simply a renunciation of imaging quality. It is a completely different piece of physics. One can say that, in contrast to imaging optics, it is physics.

Friedrich Herrmann

12.4 Radiance

Subject:

If we look through a small pipe, whose inner walls are blackened, at a monochrome, uniformly illuminated wall, we cannot decide how far away from the wall we are on the basis of what we see.

Deficiencies:

The experiment described above shows it most clearly, but the phenomenon also manifests itself without this experimental effort. It is omnipresent. One can also formulate it this way: The perceived brightness of an object does not change with distance. It is perceived by our eyes, but also by every camera. Eye and camera are good measuring instruments for it. But for what? It must be a local quantity, because it is "measured" at the location of the eye or the camera, and not at the location of the surface from which the light comes. It is a physical quantity that is not mentioned at all in school physics text books, and which is also rarely found in university physics. It is the radiance *L*.

Without this physical quantity it is also difficult to understand why it is not possible to concentrate sunlight with the help of lenses or mirrors in such a way that a temperature is reached that is higher than that of the surface of the sun. (The simple rules of geometric optics would allow such a concentration.)

Origin:

The content of our physics lessons is largely based on convention. And convention tells us to treat light with the tools of geometrical optics. The question of the distribution of the energy and that of energy currents is not addressed.

In addition, the beautiful subject becomes somewhat repulsive by the context, in which we might get to know about it: If one has worked through the many terms and definitions of photometry, as well as radiometry, one has probably lost the hope that there remains something fundamental and interesting to understand.

Disposal:

Radiance is the energy flux density per solid angle. (We are not interested in the wavelength dependence here. So we can assume that we are talking about monochromatic light.) It is a scalar quantity used to describe a radiation field locally. It depends not only on the position (*x, y, z*) within the radiation field, but also on the direction at each point (ϑ, ϕ).

This sounds complicated, but it is not. The best way to understand the quantity is to look at a radiance meter, see Figure 1.

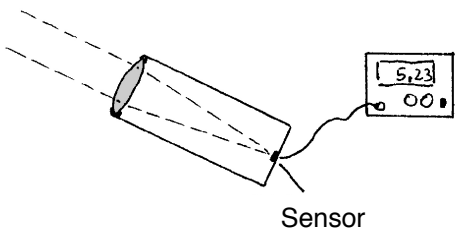


Fig. 1. Radiance meter. It measures the radiance at the position of the lens and in the direction of the optical axis of the instrument.

It measures the radiance at the position of the lens at the entrance, and for the direction of the optical axis of the device. In order to obtain the spatial distribution of the radiation, the instrument is moved around the room with the direction held constant. For the angular dependence, it is rotated in various directions at a fixed position. (If one also measures the frequency dependence, one gets the spectral radiance. It then describes the radiation field in the six-dimensional phase space.)

If one moves the device in the direction of its optical axis without changing its orientation, the measured value does not change, or more generally: The radiance in the direction of a light ray is the same at every position on the ray. The light can pass through any optical system of lenses and mirrors – the radiance does not change on a ray. Only when the light is scattered or absorbed does it change. (When entering a material with a refractive index *n*, it increases by n^2, but when exiting it returns to the initial value).

If we know this last rule, it is not difficult to understand why we can concentrate sunlight only so much as to reach the temperature of the solar surface. The best that can be achieved is that in every point of the image plane light comes from the whole half-space. Since the radiance is the same as at the start at the solar surface, the situation in the image plane is the same as directly above the solar surface: The light comes from the whole half-space. It's like holding the receiver directly in front of the sun. Thus, at best, there will be thermal equilibrium between the solar surface and the receiver.

Friedrich Herrmann

12.5 Black and white and the blue of the sky

Subject:

The color of the sky, namely blue, seems to be an important subject.

Headline in a schoolbook:
Why is the sky blue?

Headline in another schoolbook:
Scattering of light - sky blue and evening red.

Headline in a college textbook:
Why is the Sky Blue?

The color of no other system or entity is discussed in such detail.

Deficiencies:

If one asks why the subject is treated, two answers suggest themselves at first:

1. because one wants to explain the colors of things that surround us. Physics is relevant for that.

2. because one wants to deal with Rayleigh scattering, and the blue of the sky is an example that everybody knows.

Both answers do not seem very plausible to me.

To 1: With the same argument, one would have to look at other colors, or at least the most important ones: for example: Why is the wall, the snow, the cloud white? Or why is black everything that is black? Why is gold yellow or golden? But this is not done. Probably this is because it is thought that there is nothing to say: White is just when nothing is absorbed, black when everything is absorbed and yellow when blue is absorbed. By the way: If the blue of the sky is discussed, why not also the blue of the sea?

To 2: If Rayleigh scattering is considered important, so important that it belongs in a textbook, why are other scattering processes not addressed, first and foremost scattering from white objects, such as a sheet of paper or a white wall? Or is this considered trivial? But also: Why not Raman or Brillouin scattering?

One more remark on how Rayleigh scattering is explained. One talks about the role of the vibrations of the molecules, of how they act as small Hertzian antennas. But unfortunately one does so only in this context. The students learn about atomic excitations in the context of atomic physics. Are the electrons of the molecules excited here too? Into what kind of states? And if it would be so, one would not understand why the phase of the emitted coupled to that of the incident wave. Or does quantum physics no longer apply and the molecules behaves like a classical dipole antennas? The authors are honest enough to explain that the scattered light should actually interfere away, but does not do so, because density fluctuations exist in the scattering air. Please do not misunderstand me. I do not claim at all that these statements are not correct. But if the waves, which actually wanted to be scattered, interfere away – would it not have been suitable to problematize this effect where it actually takes place, namely always when the light passes through a transparent body – liquid or solid – and not only when it does not take place?

Origin:

Could it be that one only wants to express: Look, physics is not so boring as you always thought! If you plan to write a poem about the beauty of the blue sky, you should look into your physics book before, so that you know what you are talking about.

Disposal:

The interaction of light with matter can be divided into categories: absorption, refraction, reflection and scattering. In general, all processes take place simultaneously and they are wavelength dependent. Matter is complicated, and therefore these processes are also complicated and diverse. Among them, however, there are simple special cases. And one will discuss preferably these. These include in particular the origin of the body colors black and white. Their treatment in the classroom is suitable because the phenomena are universal.

The white of the cloudy sky, the snow, the white wall, the paper, the milk, the leaves of the margarite, a T-shirt or bed sheet always comes about in the same way: through many repeated refraction processes (in which practically nothing is absorbed), all the light that falls on a surface eventually gets out again. This is an interesting phenomenon, because one might expect that not every light ray would succeed in finding its way back to the surface. The process is called scattering - without determiners like Rayleigh, Mie, Raman, Brillouin, Compton, Rutherford or Thomson.

The process by which black comes about is similarly universal. The black of a black-painted wall, of a black-painted car, of the soot in a stovepipe, of printed letters comes about by substances which, if they are present as a smooth surface, reflect. Reflection, however, basically does not work without absorption (in contrast to refraction). You can see it clearly if you put two mirrors opposite each other and look into them at an angle. The light reflected several times becomes weaker and weaker. If you now grind such a material, the light initially has a similar behavior as with white bodies: it is reflected back and forth in a disordered zigzag pattern. With each reflection, however, it loses intensity, so that it runs dead in the material.

So we do not use any molecular or atomic physical interpretation for the colors black and white. By the way, the blackest black is obtained by poking a small hole in a cardboard box, such as a shoe box. The hole is black even if the inside walls of the box are white.

Friedrich Herrmann

13

Astrophysics

13.1 How the sun is working

Subject:

Luminosity, radiant power, photosphere, Fraunhofer lines, differential rotation, granulation, chromosphere, corona, flares, solar wind, solar activity, sunspots, solar prominence, chromospheric eruptions, electron, muon and tauon neutrinos, perchloroethylene, CNO cycle, p-p chain, Bethe-Weizsäcker cycle, central region, radiative transfer region, hydrogen convection zone, and other things and phenomena are introduced and discussed in specialist books on the physics of the sun. But not only there. I also found them all in textbooks for the school. Actually I had looked for something else.

Deficiencies:

Too much unimportant stuff is told, and at the same time too little about how the sun is working.

For the selection of the contents of my lessons, I have always tried to take into account the following criterion: What would I want my students to remember when they have forgotten 50%, or 90%, or even 99% of what was covered in class? When you have considered this, you also have an answer to the question: What can I say about the subject if I cannot devote to it more than one or two lessons?

For example, do I really expect my students to be able to recite the Bethe-Weizsäcker and CNO cycles? Or do they perhaps at least need to know the two names of the reactions? A comparison with chemistry is helpful: Here, no one would think of breaking down the combustion of hydrocarbons (such as octane or benzene) into a sequence of *x* intermediate reactions.

I also noticed that in connection with the sun phenomena are discussed, or even only described, which occur likewise on or in the earth, but are not addressed here. For example, details about the solar magnetism are treated. On the other hand, not a word is said about the functioning of the geodynamo. The energy transport to the surface of the sun is described: inside by scattering of photons, outside convective. The same phenomena could be discussed in the context of the earth's atmosphere, and they would indeed be a worthwhile topic there.

It is my impression that what the publishers sell as a physics textbook is more alike a work of reference. The question also arises: Does each of the 10 or 15 authors know all that the other 9 or 14 have written? After all, isn't it somewhat suspicious that physics textbooks (like chemistry textbooks) have significantly more authors than, say, French or English textbooks?

Now, in my opinion, there is not only too much told about the sun, but also too little. Here are two questions that someone who has any interest at all in the subject could certainly ask:

- If the same reaction takes place in the sun as in a hydrogen bomb, why doesn't the sun explode?
- If the Sun consists essentially of hydrogen and helium, why is its spectrum not a helium-hydrogen line spectrum?

Origin:

As is so often the case with an insufficiently elementarized subject, everything that has accumulated in the technical literature over the course of time is dumped in the classroom. Of course, the discovery of nuclear reaction cycles was important. It showed that the idea of the nuclear origin of the energy was correct. This was a big problem at that time, i.e., about 100 years ago. But now that we know, it is enough to say: Yes, it is the conversion of hydrogen into helium that provides the energy.

Disposal:

This is not the place to describe a classroom course. Instead, just a suggestion for interesting topics, roughly in order of importance.

1. Some data of the sun, including mass density as a function of the distance to the center (90% of the sun's mass is within half the radius).

2. The energy comes from a reaction of hydrogen to helium (net reaction only).

3. When heat is supplied to the sun it becomes larger and colder (i.e. not warmer). Thus a negative feedback exists so that the nuclear reaction is very slow and very stable. (One discusses the cause of this negative feedback).

4. The energy transport inside is conductive (the carrier particles are photons), outside convective. The inside is analogous to a stable stratification of the Earth's atmosphere, where the energy transport to the outside is diffusive with infrared photons. The outside corresponds to an unstable stratification of the atmosphere.

5. All matter becomes opaque if the layer is sufficiently thick (discuss cause). In the outer region of the sun this layer is about 500 km thick. This is very little compared to the diameter of the sun. Therefore the sun seems to have a sharp edge seen from us and radiates like a black body.

Friedrich Herrmann

13.2 White dwarfs, part 1: Pressure or force equilibrium?

Subject:

"Classical particle movement is not enough to balance gravity."

"In a main sequence star, the thermal energy $E_{kin} = (3/2)kT$ of the nuclei that have been stripped of all electrons generates a pressure that can withstand the gravitational pressure."

"The pressure p generated by the degenerated electron gas – $p = dE/dV$ – balances the gravitational pressure."

"The trapped electrons exert a force outwards that balances the gravitational force."

Deficiencies:

Probably everyone knows what is meant. But things get difficult when one tries to reconcile the statements with what is remembered from the mechanics classes. Let's go through one quote after the other.

The movement keeps gravity in equilibrium? In every equilibrium, whether thermal, mechanical or chemical, the values of a physical quantity have the same value in two subsystems: two temperatures, two forces,...

But here: Which quantity has twice the same value?

Next: Movement is a process, gravity rather a phenomenon. How can they balance each other? Or does gravity mean the gravitational force? Then what is the body on which it acts?

Regarding the second quote: The thermal energy produces a pressure? So one physical quantity produces another? Can a velocity also produce a temperature, or energy momentum? But there is more: the pressure thus generated withstands the gravitational pressure. But how? Does it have the same value, or perhaps the negative of the gravitational pressure? But anyway, what is meant by gravitational pressure?

In the third quote, the pressure generated by the degenerated electron gas (Whoops! Does the gas generate its own pressure? Does the air also produce the air pressure? Does it perhaps also produce its own temperature? Well, it was probably simply meant: the pressure of the degenerate electron gas) keeps the gravitational pressure in equilibrium. Yes, it seems to be meant like this: not equilibrium of forces, but equilibrium of pressures.

In the fourth quote we learn: it is indeed forces that are balancing each other. But it's not quite simple here either: The electrons exert a force outwards, i.e. every electron exerts ... a force outwards? Really? Outwards means: to the right and left, up and down, forward and backward. But that's not one force, that's at least six. The gravitational force seems to be clearer, it somehow comes from inside the star and pulls the electron downwards. The question remains, how do the six forces of the electron handle it?

So what do those for whom these texts are written do? Quite simply: In case of need (in the examination), they simply repeat the misunderstood. They have resigned themselves to the fact that they do not understand physics anyway.

Origin:

Mechanics is difficult. Apparently not only for the students.

Disposal:

Regarding the pressure equilibrium: One has a closed cylinder (Fig. 1) in which a piston can move back and forth. If the piston is not held in place, it adjusts itself so that the same pressure prevails on both sides. Pressure equilibrium is established between right and left.

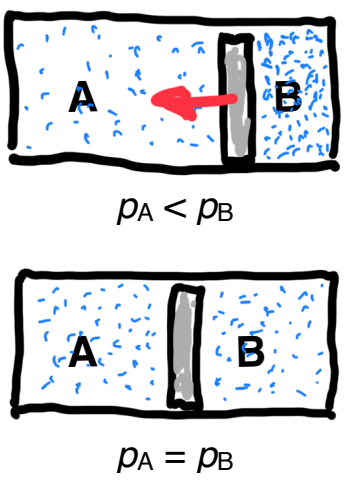


Fig. 1. Pressure equilibrium is established between subsystem A and subsystem B.

If one wants to argue with the pressure equilibrium, at least this must be known to the students. But then the problem arises: What are the two containers whose volumes can change? So it is probably better not to argue with the pressure equilibrium.

And regarding the Equilibrium of forces: there is an equilibrium of forces if two (or more) forces act on a body and add up to zero. In the present case, the forces on a portion of material in a small area of space can be considered. One of these forces is the gravitational force, the other is calculated from the pressure difference between the top and bottom of the portion of material, well-known to the students as the buoyancy force. Perhaps one could have reminded of this.

So how about the disposal? Of course, one can treat the subject correctly, but one should consider whether it is worth the effort. After all, the same discussion could have been made in the context of the interior of the earth, but of course nobody does that.

Friedrich Herrmann

13.3 White dwarfs, part 2: Rituals of explanation

Subject:

If a textbook wants to tell something, but not much, about the topic “white dwarf”, it looks something like this:

“If the mass of the star is not greater than two solar masses, a white dwarf with a radius of about 10^7 m will be formed. The further densification by gravity is counteracted by the fact that the Pauli principle forbids protons, electrons and neutrons to have an identical quantum state at the same time at the same location. Particles (in this case mainly electrons) would be forced to assume higher energy states. This circumstance counteracts further compression. One speaks of degenerate matter.”

Some textbooks go into more detail, about half a page, using the following terms, among others:

- potential well
- quantization of the energy
- Boltzmann distribution
- degenerated electron gas
- fermions
- Pauli principle
- spin orientation

Deficiencies:

1. In any case, the goal seems to be to explain why a white dwarf does not collapse. However, the reason it does not do so is the same as that the earth does not collapse or that solid or liquid materials are hard. There is usually no word about why it is difficult to compress condensed matter on the earth. This gives the impression that a white dwarf is something unusual and difficult; it cannot be understood without quantum mechanics; the physics that is effective is not the same as on Earth.

2. White dwarfs have an interesting characteristic in which they differ significantly from the objects of our earthly experience, but which does not seem to be worthy of any comment or explanation: When matter is added to a white dwarf, it does not become larger, but smaller. This has less to do with quantum physics than with ordinary classical gravitation.

Origin:

The description is taken from specialist literature or the university physics book. There the effort is necessary, because there the goal is to calculate the Chandrasekar limit. For this purpose one needs two ingredients: The equation of state and the law of gravity. Both together lead to a differential equation, the Lane-Emden equation, which is not easy to solve. Now the equation of state, which is actually quite simple because it does not contain the temperature, and therefore only describes the relationship between pressure and mass density, cannot be measured directly, because this would require pressures that cannot be produced in any laboratory. So one has to calculate them, and that requires the effort quoted.

However, the quantum mechanical paraphernalia are presented in school lessons, although the equation of state is even not mentioned.

It is an example of how school physics takes over something from specialized physics, losing sight of the learning objective. Without asking the question: Which of the topics provided by the specialist’s physics do we want to declare to be a learning objective for a general education? In this way, something is created that could aptly be described as a learning ritual.

Disposal:

The first thing to explain is that one can compress solid bodies or liquids at will if one only makes the pressure high enough - a fact that is not really surprising. Sufficiently high pressures come up if a compact celestial body is made heavier and heavier. This happens with some white dwarfs: they gradually receive matter from a partner star, with which they form a double star system.

Our normal experience is: If one pours more sand on a pile of sand, the pile will get higher. The white dwarf behaves differently. When matter is added, the matter sinks. It approaches the center, and as it does so the gravitational field strength increases, and with it the gravitational force. It increases so much that the whole star shrinks. This cannot happen in a sand pile, because the gravitational force comes practically only from the Earth. The sand pile and its size does not matter to them, because it is small against the Earth as a whole.

Figure 1 shows how the radius of a white dwarf decreases with increasing mass, and how the star eventually even collapses. The collapse ends when the atomic nuclei touch each other.

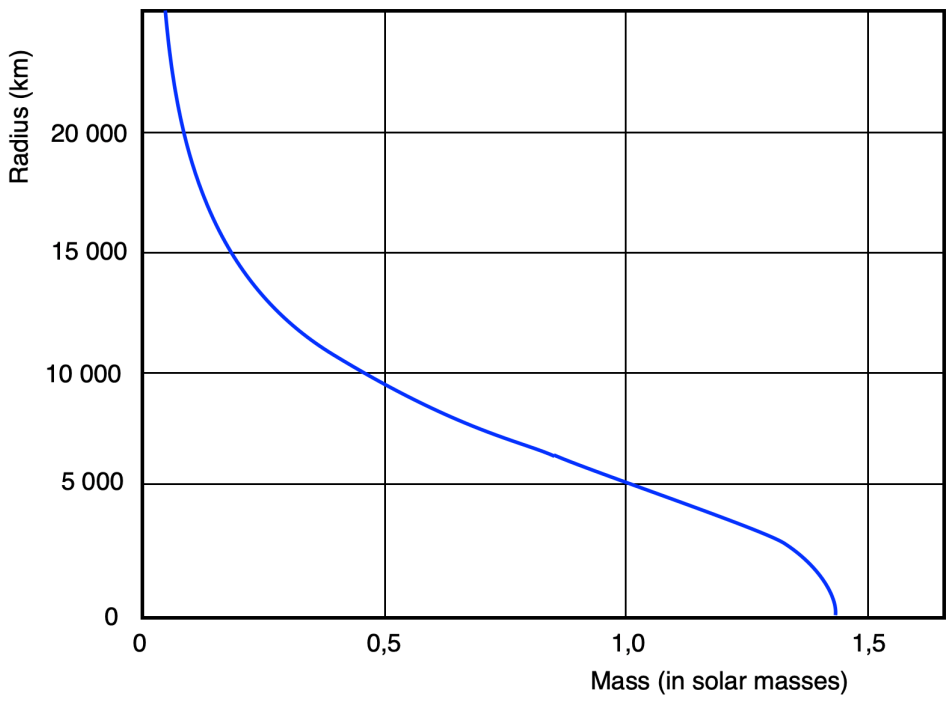


Fig. 1. Radius of a white dwarf as a function of its mass

It can also be said that when the sunlike star becomes a white dwarf, the atoms condense. When the white dwarf becomes a neutron star, the nuclear matter condenses. The nuclear matter is much harder than that of the white dwarf (and the Earth).

It can also be mentioned that the behaviour is determined in detail by the equation of state $\rho(p)$, i.e. an extension of Hooke’s law. But all this please without Pauli, Fermi & Co.

Friedrich Herrmann