Irreversibility and the second law of thermodynamics

Jos Uffink *

July 5, 2001

1 INTRODUCTION

The second law of thermodynamics has a curious status. Many modern physicists regard it as an obsolete relic from a bygone age, while many others, even today, consider it one of the most firmly established and secure achievements of science ever accomplished.

From the perspective of the foundations of physics, a particularly interesting question is its relationship with the notion of irreversibility. It has often been argued, in particular by Planck, that the second law expresses, and characterises, the irreversibility of all natural processes. This has led to much debated issues, such as whether the distinction between past and future can be grounded in the second law, or how to reconcile the second law with an underlying microscopic mechanical (and hence reversible) theory. But it is not easy to make sense of these debates since many authors mean different things by the same terms.

The purpose of this paper is to provide some clarification by distinguishing three different meanings of the notion of (ir)reversibility and to study how they relate to different versions of the second law. A more extensive discussion is given in [23].

2 THREE CONCEPTS OF (IR)REVERSIBILITY

Many physical theories employ a state space Γ consisting of all possible states of a system. An instantaneous state is thus represented as a point s in Γ and a process

*Institute for History and Foundations of Science, PO Box 80.000 3508 TA Utrecht, the Netherlands uffink@phys.uu.nl

as a parameterised curve:

$$\mathcal{P} = \{s_t \in \Gamma \,:\, t_i \leq t \leq t_f\}.$$

The laws of a theory usually allow only a definite class of such processes (e.g. the solutions of the equations of motion). Call this class $\mathcal{W}$, the set of all possible worlds (according to this theory). Let now R be an involution (i.e. $R^2 s = s$) that turns a state s into its 'time reversal' Rs. In classical mechanics, for example, R is the transformation which reverses the sign of all momenta and magnetic fields. In a theory like classical thermodynamics, in which the state does not contain velocity-like parameters, one may take R to be the identity transformation. Further, define the time reversal $\mathcal{P}^*$ of a process $\mathcal{P}$ by:

$$\mathcal{P}^* = \{(Rs)_{-t} \,:\, -t_f \leq t \leq -t_i\}.$$

The theory (or a law) is called *time-reversal invariant* (TRI) if the class $\mathcal{W}$ is closed under time reversal, i.e. iff:

$$\mathcal{P} \in \mathcal{W} \implies \mathcal{P}^* \in \mathcal{W}. \tag{1}$$

According to this definition[1] the form of the laws themselves (and a given choice for R) determines whether the theory is TRI or not. And it is straightforward to show that classical mechanics is indeed TRI. Note also that the term 'time-reversal' is not meant literally. That is to say, we consider *processes* whose reversal is or is not allowed by a physical law, not a reversal of time itself. The prefix is only intended to distinguish the term from a spatial reversal. Furthermore, note that it is not relevant here whether the reversed processes $\mathcal{P}^*$ occur in the actual world. It is sufficient that the theory allows them. Thus, the fact that the sun never rises in the west is no obstacle to celestial mechanics qualifying as TRI.

Is this theme of time-reversal (non)invariance related to the second law? Even though the criterion is unambiguous, its application to thermodynamics is not a matter of routine. In contrast to mechanics, thermodynamics does not possess equations of motion. This, in turn, is due to the fact that thermodynamical processes only take place after an external intervention on the system. (E.g.: removing a partition, establishing thermal contact with a heat bath, pushing a piston, etc.) They do not refer to the autonomous behaviour of a free system. This is not to say that time plays no role. Classical thermodynamics, in the formulation of

[1]Some authors [14] propose an alternative definition of time reversal invariance. Supposing the theory is deterministic, its laws specify evolution operators $U(t_1, t_0)$ such that $s_t = U(t, t_0)s_0$. In that case, one can define TRI of the theory by the requirement $U^{-1}(t, t_0)RU(t, t_0) = R$. This definition has the advantage that it does not rely on the possible worlds semantics. However, it applies only for deterministic theories, and not to thermodynamics.

Clausius, Kelvin or Planck, is concerned with processes occurring in the course of time, and its second law is clearly not TRI. However, in other formulations, such as those by Gibbs, Carathéodory or Lieb and Yngvason, this is less clear.

My main theme, however, is the notion of '*(ir)reversibility*'. This term is attributed to processes rather than theories or laws. But in the philosophy of physics literature it is intimately connected with time-reversal invariance. More precisely, one calls a process $\mathcal{P}$ allowed by a given theory irreversible iff the reversed process $\mathcal{P}^*$ is excluded by this theory. Obviously, such a process $\mathcal{P}$ exists only if the theory in question is not TRI. Conversely, every non-TRI theory admits irreversible processes in this sense. They constitute the hallmark of time-reversal variance and, therefore, discussions about (ir)reversibility and (non)-TRI in philosophy of physics coincide for the most part. However, in thermodynamics, the term is commonly employed with other meanings.

The thermodynamics literature often uses the term 'irreversibility' to denote an aspect of our experience which, for want of a better word, one might also call *irrecoverability*. In many processes, the transition from an initial state s_i to a final state s_f, cannot be fully 'undone', once the process has taken place. In other words, there is no process which starts off from state s_f and restores the initial state s_i completely. Ageing and dying, wear and tear, erosion and corruption are the obvious examples. This is the sense of irreversibility that Planck intended, when he called it the essence of the second law.

Many writers have emphasised this theme of irrecoverability in connection with the second law. Indeed, Eddington introduced his famous phrase of 'the arrow of time' in a general discussion of the 'running-down of the universe', and illustrated it with many examples of processes involving 'irrevocable changes', including the example of Humpty-Dumpty who, allegedly, could not be put together again after his great fall. In retrospect, one might perhaps say that a better expression for this theme is the *ravages* of time rather than its arrow.

(Ir)recoverability differs from (non)-TRI in at least two respects. First, the only thing that matters here is the retrieval of the original state s_i. It is not necessary that one can find a process P^* which retraces all the intermediate stages of the original process in reverse order. A second difference is that we are dealing with a *complete* recovery. Planck repeatedly emphasised that this condition includes the demand that all auxiliary systems that may have been employed in the original process are brought back to their initial state. Now, although one might argue that a similar demand should also be included in the definition of TRI, the problem is here that the auxiliary systems are often not characterisable as thermodynamical systems.

Schematically, the idea can be expressed as follows. Let s be a state of the system and Z a (formal) state of its environment. Let $\mathcal{P}$ be some process that

brings about the following transition:

$$\langle s_i, Z_i \rangle \xrightarrow{\mathcal{P}} \langle s_f, Z_f \rangle \tag{2}$$

Then $\mathcal{P}$ is reversible in Planck's sense iff there exists[2] another process $\mathcal{P}'$ that produces

$$\langle s_f, Z_f \rangle \xrightarrow{\mathcal{P}'} \langle s_i, Z_i \rangle. \tag{3}$$

However, the term 'reversible' is also used in yet a third sense, which has no straightforward connection with the arrow of time at all. It is often used to denote processes which proceed so delicately and slowly that the system can be regarded as remaining in equilibrium 'up to a negligible error' during the entire process. We shall see that this is the meaning embraced by Clausius. Actually, it appears to be the most common usage of the term in the physical-chemical literature; see e.g. [13, 8]. A more apt name for this kind of processes is *quasi-static*. Of course, the above way of speaking is vague, and has to be amended by criteria specifying what kind of 'errors' are intended and when they are 'small'. These criteria take the form of a limiting procedure so that, strictly speaking, reversibility is here not an attribute of a particular process but of a series of processes.

Again, quasi-static processes are not necessarily the same as those called reversible in the previous two senses. For example, the motion of an ideal harmonic oscillator is reversible in the sense of Planck, but it is not quasi-static. Conversely, the discharge of a charged condenser through a very high resistance can be made to proceed quasi-statically, but even then it remains irreversible in Planck's sense.

Comparison with the notion of TRI is hampered by the fact that 'quasi-static' is not strictly a property of a process. Perhaps the following example might be helpful. Consider a process $\mathcal{P}_N$ in which a system, originally at temperature θ_1 is consecutively placed in thermal contact with a sequence of N heat baths, each at a slightly higher temperature than the previous one, until it reaches a temperature θ_2. By making N large, and the temperature steps small, such a process becomes quasi-static, and we can represent it by a curve in the space of equilibrium states. However, for any N, the time-reversal of the process is impossible.

The reason why so many authors nevertheless call such a curve 'reversible' is that one can consider a second process $\mathcal{Q}_N$, in which the system, originally at temperature θ_2, is placed into contact with a series of heat baths, each slightly *colder* than the previous one. Again, each process $\mathcal{Q}_N$ is non-TRI. *A forteriori*, no $\mathcal{Q}_N$ is the time reversal of any $\mathcal{P}_N$. Yet, if we now take the quasi-static limit, the state change of the system will follow the same curve in equilibrium space as in the previous case, traversed in the opposite direction. The point is, of course, that

[2]One might read 'exists' here as 'allowed by the theory', i.e. as $\mathcal{P}' \in \mathcal{W}$. But this is not Planck's view. He emphasised that $\mathcal{P}'$ might employ any appliances *available in Nature*, rather than allowed by a theory. This is a third respect in which his sense of reversibility differs from that of TRI.

precisely because this curve is not itself not a process, the notion of time reversal does not apply to it.

3 EARLY FORMULATIONS OF THE SECOND LAW

The work of the founding fathers Carnot, Clausius and Kelvin (=W.Thomson) can be divided into two lines: one main line dealing exclusively with cyclic processes; and another addressing (also) non-cyclic processes. In this section I will discuss both.

The first line starts with the work of Carnot (1824). Carnot studied cyclic processes performed by an arbitrary system in interaction with two heat reservoirs, (the furnace and the refrigerator), at temperatures θ^+ and θ^-, while doing work on some third body. Let $Q^+(\mathcal{C})$, $Q^-(\mathcal{C})$ and $W(\mathcal{C})$ denote, respectively, the heat absorbed from the furnace, the heat given off to the refrigerator, and the work done by the system during the cycle $\mathcal{C}$. He assumed that the heat reservoirs remain unchanged while they exchange heat with the system.

Carnot's main assumptions were: (i) heat is a conserved substance, i.e., $Q^+(\mathcal{C}) = Q^-(\mathcal{C})$; and (ii) the impossibility of a perpetuum mobile of the first kind, or:

> CARNOT'S PRINCIPLE: It is impossible to perform a repeatable cycle in which the only result is the performance of (positive) work.

Note that Carnot did not object to the performance of (positive) work in a cycle as such. Rather, his point was that, due to the assumption that the heat reservoirs act as invariable buffers, the cycle could be repeated arbitrarily often. Thus, violation of the above principle would provide *unlimited* production of work at no cost whatsoever. This he regards as inadmissible.

By a well known reductio ad absurdum argument, he obtained

> CARNOT'S THEOREM: (a) The efficiency $\eta(\mathcal{C}) := W(\mathcal{C})/Q^+(\mathcal{C})$ is bounded by a universal maximum C, which depends only on the temperatures θ^+ and θ^-:
>
> $$\eta(\mathcal{C}) \leq C(\theta^+, \theta^-) \tag{4}$$
>
> (b) This maximum is attained if the cycle $\mathcal{C}$ is 'reversible'.

In fact Carnot did not use the term 'reversible'. So one might ask how he conceived of the condition in (b). Actually he discusses the issue twice. He starts his argument with an example: the Carnot cycle for steam. In passing, he notes its relevant feature: 'The operations we have just described might have been performed in an inverse direction and order (Mendoza, 1960, p.11).' This feature, of course, turns out to be crucial for the claim that this cycle has maximal efficiency. Later, he realised that a more precise formulation of this claim was desirable, and he formulates a necessary and sufficient criterion for maximum efficiency

(ibid. p. 13): it should be avoided that bodies of different temperature come into direct thermal contact. He notes that this criterion cannot be met exactly, but can be approximated as closely as we wish. In modern terms: the criterion is that the process should be quasi-static at all stages which involve heat exchange.

Accordingly, even at this early stage, there are two plausible options for a definition of a 'reversible' cycle. Either we focus on the crucial property of the Carnot cycle that it can also be run backwards. This is the option chosen by Kelvin in 1851. Of course, this is a natural choice, since this property is essential to the proof of the theorem. Or else, one can focus on Carnot's necessary and sufficient condition and use this as a definition of a reversible cycle. This is more or less the option followed by Clausius in 1864. He called a cyclic process reversible (*umkehrbar*) iff it proceeds quasi-statically.

Carnot's work proved very valuable, a quarter of a century later, when Kelvin showed in 1848 that it could be used to devise a absolute temperature scale. But in the meantime, serious doubts had appeared about the conservation of heat. Thus, when the importance of his theorem was recognised, the adequacy of Carnot's original derivation had already become suspect. Therefore, Clausius (1850) and Kelvin (1851) sought to obtain Carnot's theorem on a different footing. They replaced Carnot's assumption (i) by the Joule-Mayer principle stating the equivalence of heat and work, i.e.: $Q^+(\mathcal{C}) = Q^-(\mathcal{C})) + JQ(\mathcal{C})$ where J is Joule's constant. Instead of Carnot's (ii), they adopted the impossibility of perpetuum mobile of the second kind:

> THE CLAUSIUS/KELVIN PRINCIPLE It is impossible to perform a cycle[3] in which the only effect is:
>
> > to let heat pass from a cooler to a hotter body (Clausius)
> > to perform work and cool a single heat reservoir (Kelvin).

They showed that Carnot's theorem, by that time called the "second thermodynamic law" or the "*Zweite Hauptsatz*" can be recovered.

In a series of papers, Clausius and Kelvin extended and reformulated the result. In 1854 Kelvin showed that the absolute temperature scale $T(\theta)$ can be chosen such that $C(T^+, T^-) = J(1 - T^-/T^+)$ or equivalently

$$\frac{Q^+(\mathcal{C})}{T^+} = \frac{Q^-(\mathcal{C})}{T^-}. \tag{5}$$

[3]In the usual formulation of these principles, the unlimited repeatability of the cycle is not stressed so much as it was by Carnot. However, one may infer that it was at least intended by Kelvin, when he wrote in his introduction 'Whenever in what follows, *the work done or the mechanical effect produced* by a thermo-dynamic engine is mentioned without qualification, it must be understood that the mechanical effect produced, either in a non-varying machine, or in a complete cycle, or any number of complete cycles of a periodical engine, is meant.' [15, p. 177].

Generalising the approach to cycles involving an arbitrary number of heat reservoirs, they obtained the formulation[4]

$$\oint_C \frac{đQ}{T} = 0 \quad \text{if } C \text{ is reversible,} \tag{6}$$

and

$$\oint_C \frac{đQ}{T} \leq 0 \quad \text{if } C \text{ is not reversible} \tag{7}$$

Note that here T stands for the absolute temperature of the heat reservoirs; it is only in the case of (6) that T can be equated with the temperature of the system.

Let us now investigate the connections with the themes of section 2. All three authors adopt a principle which is manifestly non-TRI: it forbids the occurrence of certain cyclic processes while allowing their reversal. However, the main objective in the work of Clausius and Kelvin considered above was to obtain part (a) of Carnot's theorem, or its generalisation (6). These results are TRI, and accordingly, the non-TRI element did not receive much attention. Indeed, Kelvin never mentions relation (7) at all, and indeed calls (6) "the full expression of the second thermodynamic law". Clausius (1854) discusses (7) only very briefly.

It is much harder to find a connection with irrecoverability. All the papers considered here are only concerned with cyclic processes. There can be no question, therefore, of irrecoverable changes in that system, or of a monotonically changing quantity. If one insists on finding such a connection, the only option is to take the environment into account, in particular the heat reservoirs. Indeed, nowadays one would argue that if a system performs an irreversible cycle, the total entropy of the heat reservoirs increases. But such a view would be problematic here. First, we are at a stage in which the very existence of an entropy function is yet to be established. One cannot assume that the heat reservoirs already possess an entropy, without running the risk of circularity. Moreover, the heat reservoirs are conceived of as buffers with infinite heat capacity, and it is not straightforward to include them in an entropy balance. The connection with irrecoverability therefore remains dubious.

The second line announced at the beginning of this section consists mainly of three papers: Kelvin (1852) and Clausius (1864) and (1865). They differ from earlier and later works of the same authors because they explicitly address non-cyclic processes. Kelvin (1852) is a very brief note on the 'universal tendency towards the dissipation of energy'. He argued that natural processes in general bring about 'unreversible' changes, so that a full restoration of the initial state is impossible. Clearly, Kelvin uses the term 'unreversible' here in the sense of 'irrecoverable'. He claims that this tendency is a necessary consequence of his

[4]From here on, Q is regarded as positive when absorbed by, and negative when given off by the system.

(1851)principle mentioned above. Moreover, he draws an eschatological conclusion: in the distant future, life on earth must perish. It is here that we first encounter the "terroristic nimbus" of the second law: the heat death of the universe.[5]

Starting in 1862, Clausius also addresses non-cyclic processes, and some years later, reaches a similar conclusion. He notes in 1865 that the validity of (6) implies that the integral $\int_{s_1}^{s_2} \frac{\bar{d}Q}{T}$ is independent of the integration path, and can be used to define a new function of state, called entropy S, such that

$$S(s_2) - S(s_1) = \int_{s_1}^{s_2} \frac{đQ}{T} \tag{8}$$

where the integral is performed for an *umkehrbar* (i.e. quasi-static) process. For an *unumkehrbar* process he uses relation (7) to obtain

$$\int_{s_1}^{s_2} \frac{đQ}{T} \leq S(s_2) - S(s_1). \tag{9}$$

If this latter process is adiabatic, i.e. if there is no heat exchange with the environment, one may put $đQ = 0$ and it follows that

$$S(s_2) \geq S(s_1). \tag{10}$$

Hence we obtain:

> THE SECOND LAW (Clausius' version) For every non-quasi-static process in an adiabatically isolated system which begins and end in an equilibrium state, the entropy of the final state greater than or equal to that of the initial state. For every quasi-static process in an adiabatical system, the entropy of the final state is equal to that of the initial state.

This is the first instance of a formulation of the second law as a statement about entropy increase. Note that only the '$\geq$' sign is established. One often reads that for irreversible processes the strict inequality holds in (10), holds but this does not follow from Clausius' version. Note also that, in contrast to the common view that the entropy principle holds for isolated systems only, Clausius' result applies to *adiabatically* isolated systems.

Clausius too draws a bold inference about all natural processes and the fate of the universe:

> 'The second law in the form I have given it says that all transformations taking place in nature go by themselves in a certain direction,

[5]The lack of any argument for Kelvin's bold claims has puzzled many commentators. It has been suggested [**?**] the source for these claims is perhaps to be found in his religious beliefs rather than in thermodynamics.

> which I have denominated the positive direction. [...] The application of this law to the universe leads to a conclusion to which W. Thomson first called attention [...] namely, [...] that the total state of the universe will change continually in that direction and hence will inevitably approach a limiting state.' [6, p. 42]

Noting that his theory is still not capable of treating the phenomenon of heat radiation, he 'restricts himself' —as he puts it— to an application of the theory to the universe:

> 'One can express the fundamental laws of the universe that correspond to the two main laws of thermodynamics in the following simple form:
> 1. The energy of the universe is constant.
> 2. The entropy of the universe tends to a maximum.' (ibid. p. 44)

These words of Clausius are probably the most often quoted, and the most controversial,in the history of thermodynamics. Even Planck admitted that the entropy of the universe is an undefined concept [18, § 135]. Ironically, Clausius could have avoided such criticism if he had not 'restricted' himself to the universe but generalised his formulation to an arbitrary adiabatically isolated system (beginning and ending in equilibrium).

Another objection is that this version of the Law presupposes that the initial and final states can also be connected by a quasi-static process, in order to define their entropy difference by means of (8). This is not trivial for transformations other than exchanges of heat and work.

To conclude, this second line of development focuses on arbitrary non-cyclic processes of completely general systems. The main claim is that, apart from the quasi-static case, all such processes are irrecoverable. However, the arguments given for those grand claims are rather fragile. Kelvin provides no argument at all, and Clausius' attempts depends on rather special assumptions. A curious point is that when Clausius reworked his previous papers into a textbook (1876) he completely dropped his famous claim that the entropy of the universe tends to a maximum. The most general statement of the second law presented in this book, is again given as (6) and (7), i.e. restricted to cyclic processes.

4 PLANCK

The importance of Planck's *Vorlesungen über Thermodynamik*[18] can hardly be underestimated. The book has gone through eleven editions, from 1897 until 1964, and still remains the most authoritative exposition of classical thermodynamics. Planck's position has always been that the second law expresses irrecoverability of all processes in nature. However, it is not easy to analyse Planck's arguments for this claim. His text differs in many small but decisive details in the various

editions. I also warn that the English translation of the *Vorlesungen* is unreliable. Particularly confusing is that it uses the translation 'reversible' indiscriminately, where Planck distinguishes between the terms *umkehrbar*, which he uses in Clausius' sense, i.e. meaning 'quasi-static', and *reversibel*, in the sense of Kelvin (1852) meaning 'recoverable'. Moreover, after Planck learned about Carathéodory's work through a review by Born in 1921, he presented a completely different argument from the eighth edition onwards.

In spite of the many intricacies in Planck's book, I shall limit myself to a brief exposition of Planck's latter argument, published first in [19]. He starts from the statement that "friction is an *irreversibel* process", which he considers to be an expression of Kelvin's principle. This may need some explanation, because, at first sight, this statement does not concern cyclic processes or the *perpetuum mobile* at all. But for Planck, the statement means that there exists no process which 'undoes' the consequences of friction, i.e., a process which produces no other effect than cooling a reservoir and doing work. The condition 'no other effect' here allows for the operation of any type of auxiliary system that operates in a cycle.

He then considers an adiabatically isolated fluid[6] capable of exchanging energy with its environment by means of a weight at height h. Planck asks whether it is possible to bring about a transition from an initial state s of this system to a final state s', in a process which brings about no changes in the environment other than the displacement of the weight. If Z denotes the state of the environment and h the height of the weight, the desired transition can be represented as

$$(s, Z, h) \xrightarrow{?} (s', Z, h').$$

He argues that, by means of '*reversibel*-adiabatic'[7] processes, one can always achieve a transition from the initial state s to an intermediary state s^* in which the volume equals that of state s' and the entropy equals that of s. That is, one can realise a transition

$$(s, Z, h) \longrightarrow (s^*, Z, h^*), \quad \text{with} \quad V(s^*) = V(s') \ \text{ and } \ S(s^*) = S(s).$$

Whether the desired final state s' can now be reached from the intermediate state s^* depends on the value of the only independent variable in which s^* and s' differ. For this variable one can choose the energy U.

There are three cases:
(1) $h^* = h'$. In this case, energy conservation implies $U(s^*) = U(s')$. Because the coordinates U and V determine the state of the fluid completely, s^* and s' must coincide.

[6] A fluid has, by definition, a state completely characterised by two independent variables.
[7] Apparently, Planck's pen slipped here. He means: *umkehrbar*-adiabatic.

(2) $h^* > h'$. In this case, $U(s^*) < U(s)$, and the state s' can be reached from s^* by letting the weight perform work on the system, e.g. by means of friction, until the weight has dropped to height h'. According to the above formulation of Kelvin's principle, this process is irreversible (i.e. irrecoverable).
(3) $h^* < h'$ and $U(s^*) > U(s)$. In this case the desired transition is impossible. It would be the reversal of the irreversible process just mentioned in (2), i.e. produce work by cooling the system and thus realise a *perpetuum mobile* of the second kind.

Now, Planck argues that in all three cases, one can also achieve a transition from s^* to s' by means of heat exchange in an *umkehrbar* (i.e. quasi-static) process in which the volume remains fixed. For such a process he writes

$$dU = TdS. \tag{11}$$

Using the assumption that $T > 0$, it follows that, in the three cases above, U must vary in the same sense as S. That is, the cases $U(s^*) < U(s')$, $U(s^*) = U(s')$ or $U(s^*) > U(s')$, can also be characterised as $S(s^*) < S(s')$, $S(s^*) = S(s')$ and $S(s^*) > S(s')$ respectively.

An analogous argument can be constructed for a system consisting of several fluids. Just as in earlier editions of his book, Planck generalises the conclusion (without a shred of proof) to arbitrary systems and arbitrary physical/chemical processes:

> Every process occurring in nature proceeds in the sense in which the sum of the entropies of all bodies taking part in the process is increased. In the limiting case, for reversible processes this sum remains unchanged. [...] This provides an exhaustive formulation of the content of the second law of thermodynamics [19, p. 463]

Note how much Planck's construal of the *perpetuum mobile* differs from Carnot and Kelvin. The latter authors considered the device which performs the cycle, as the system of interest and the reservoir as part of the environment. By contrast, for Planck, the *reservoir* is the thermodynamical system, and the engine performing the cyclic process belongs to the environment. Related to this switch of perspective is the point that the reservoir is now assumed to have a finite energy content. Thus, the state of the reservoir can change under the action of the hypothetical *perpetuum mobile* device. As a consequence, the cycle need not be repeatable, in sharp contrast to Carnot's original formulation of the idea.

Secondly, Planck's argument can hardly be regarded as satisfactory for the bold and universal formulation of the second law. It applies only to systems consisting of fluids, and relies on several implicit assumptions which can be questioned outside of this context. In particular, this holds for the assumption that there always exist functions S and T (with $T > 0$) such that $đQ = TdS$; and the assumption of a rather generous supply of quasi-static processes. As we shall see in section 6, Carathéodory's treatment is much more explicit on just these issues.

5 GIBBS

The work of Gibbs [11] is very different from that of his European colleagues. Where they were primarily concerned with processes, Gibbs concentrates his efforts on a description of equilibrium states. He assumes that these states are completely characterised by a finite number of state variables like temperature, energy, pressure, volume, entropy, etc., but he makes no effort to prove the existence or uniqueness of these quantities from empirical principles. He proposes:

> THE PRINCIPLE OF GIBBS: For the equilibrium of any isolated system it is necessary and sufficient that in all possible variations of the state of the system which do not alter its energy, the variation of its entropy shall either vanish or be negative. [11, p.56]

Actually, Gibbs presented this statement only as "an inference naturally suggested by the general increase of entropy which accompanies the changes occurring in any isolated material system". But many later authors have regarded the Gibbs principle as a formulation of the second law (e.g. [24],[3] and [2]). We can follow their lead and Truesdell [22] about how the principle is to be understood.

The first point to note is then that the Gibbs principle is not literally to be seen as a criterion for equilibrium. Indeed, this would make no sense because all states considered here are already equilibrium states. Rather, it is to be understood as a criterion for *stable* equilibrium. Second, the principle is interpreted analogous to other well-known variational principles in physics like the principle of least action, etc. Here, a 'variation' is *virtual*, i.e., it represents a comparison between two conceivable models or 'possible worlds', and one should not think of them as (part of) a process that proceeds in the course of time in one particular world. Instead, a variational principle serves to decide which of these possible worlds is physically admissible, or, in the present case, stable.

According to this view, the Gibbs principle tells us when a conceivable equilibrium state is stable. Such a proposition obviously has a modest scope. First, not all equilibrium states found in Nature are necessarily stable. Secondly, Gibbs' principle is more restricted than Clausius' statement of the second law in the sense that it applies to a *isolated* (i.e. no energy exchange is allowed) and not merely adiabatically isolated systems. More importantly, it provides no information about evolutions in the course of time; and a direction of natural processes, or a tendency towards increasing entropy, cannot be obtained from it.Hence, the second law as formulated by Gibbs has no connections with the arrow of time.

Of course the view sketched above does not completely coincide with Gibbs' own statements. In some passages he clearly connects virtual variations to actual processes, e.g. when writing: "it must be regarded as generally possible to produce that variation by some process"[11, p. 61]. Some sort of connection between variations and processes is of course indispensable if one wants to maintain the idea that this principle has implications for processes.

Probably the most elaborate attempt to provide such a connection is the presentation by Callen [3]. Here, it is assumed that, apart from its actual state, a thermodynamic system is characterised by a number of *constraints*, determined by a macroscopic experimental context. These constraints single out a particular subset C of Γ, consisting of states which are consistent with the constraints. It is postulated that in stable equilibrium, the entropy is maximal over all states in C.

A process is then conceived of as being triggered by the cancellation of one or more of these constraints. (E.g.: mixing or expansion of gases after the removal of a partition, loosening a previously fixed piston, etc.) It is assumed that a such a process sets in spontaneously, after the removal of a constraint.

Now, clearly, the removal of a constraint implies an enlargement of the set C. Hence, if we assume that the final state of this process is again a stable equilibrium state, it follows immediately that every process ends in a state of higher (or at best equal) entropy.

I will not attempt to dissect the problems that Callen's approach brings along, except for three remarks. First, the idea of extending the description of a thermodynamical system in such a way that apart from its state, it is also characterised by a constraint brings some conceptual difficulties. For if the actual state is s, it is hard to see how the class of other states contained in the same constraint set C is relevant to the system. It seems that on this approach the state of a system does not provide a complete description of its thermodynamical properties.

Second, the picture emerging from Callen's approach is somewhat anthropomorphic. For example he writes, for the case that there are no constraints, i.e. $C = \Gamma$, that 'the system is free to select any one of a number of states' (1960, p. 27). This sounds as if the system is somehow able to 'probe' the set C and chooses its own state from the options allowed by the constraints.

Third, the result that entropy increases in a process from one equilibrium state to another, depends rather crucially on the assumption that processes can be successfully modelled as the removal of constraints. But, clearly, this assumption does not apply to all natural processes. For instance, one can also trigger a process by imposing additional constraints. Hence, this approach does not attain the universal validity of the second law that Planck argued for.

6 Carathéodory

Carathéodory [4] was the first mathematician to pursue a rigorous formalisation of the second law. Like Gibbs, he construed thermodynamics as a theory of equilibrium states rather than (cyclic) processes. Again, a thermodynamical system is described by a state space Γ, represented as a (subset of a) n-dimensional manifold in which the thermodynamic state variables serve as coordinates. He assumes that Γ is equipped with the standard Euclidean topology. But metrical properties of the space do not play a role in the theory, and there is no preference

for a particular system of coordinates.

However, the coordinates are not completely arbitrary. Carathéodory distinguishes between 'thermal coordinates' and 'deformation coordinates'. (In typical applications, temperature or energy are thermal coordinates, whereas volumes are deformation coordinates.) The *state* of a system is specified by both types of coordinates; the *shape* (*Gestalt*) of the system by the deformation coordinates alone. It seems to be assumed that the deformation coordinates remain meaningful in the description of the system when the system is not in equilibrium, whereas the thermal coordinates are defined only for equilibrium states. In any case, it is assumed that one can obtain every desired final shape from every initial state by means of an adiabatic process.

The idea is now to develop the theory in such a way that the second law provides a characteristic mathematical structure of state space. The fundamental concept is a relation that represents whether state t can be reached from state s in an adiabatic process.[8] This relation is called *adiabatic accessibility*, and I will denote it, following Lieb and Yngvason, by $s \prec t$. This notation may suggest that the relation has the properties of an ordering relation. And indeed, given its intended interpretation, this would be very natural. But Carathéodory does not state or rely on these properties anywhere in his paper.

In order to introduce the second law, Carathéodory proposes an empirical claim: from an arbitrary given initial state it is not possible to reach every final state by means of adiabatic processes. Moreover, such inaccessible final states can be found in every neighbourhood of the initial state. However, he immediately rejects this preliminary formulation, because it fails to take into account the finite precision of physical experiments. Therefore, he strengthens the claim by the idea that there must be a small region surrounding the inaccessible state, consisting of points which are also inaccessible.

The second law thus receives the following formulation:

> THE PRINCIPLE OF CARATHÉODORY: In every open neighborhood $U_s \subset \Gamma$ of an arbitrary chosen state s there are states t such that for some open neighborhood U_t of t: all states r within U_t cannot be reached adiabatically from s. Formally:
>
> $$\forall s \in \Gamma \, \forall U_s \, \exists t \in U_s \; \& \; \exists U_t \subset U_s \, \forall r \in U_t \; : \; s \not\prec r. \tag{12}$$

He then restricts his discussion to so-called 'simple systems', defined by four additional conditions:

[8] Carathéodory's definition of an 'adiabatic process' is as follows. He calls a container adiabatic if the system contained in it remains in equilibrium, regardless of what occurs in the environment, as long as the container is not moved nor changes its shape. Thus, the only way of inducing a process of a system in an adiabatic container is by deformation of the walls of the vessel. (E.g. a change of volume or stirring.) A process is said to be adiabatic if it takes place while the system remains in an adiabatic container.

1. The system has only a single independent thermal coordinate. Physically, this means that the system has no internal adiabatic walls since in that case it would have parts with several independent temperatures. By convention, the state of a simple system is written as $s = (x_0, \ldots, x_{n-1})$, where x_0 is the thermal coordinate.

2. For any given pair of an initial state and final shape of the system there is more than one adiabatic process $\mathcal{P}$ that connects them, requiring different amounts of work. For example, for a gas initially in any given state one can obtain an arbitrary final value for its volume by adiabatic expansion or compression. This change of volume can proceed slowly or fast, and these procedures indeed require different amounts of work.

3. The amounts of work done in the processes just mentioned form a connected interval. In other words, if for a given initial state and final shape there are adiabatic processes $\mathcal{P}_1, \mathcal{P}_2$ connecting them, which require the work $W(\mathcal{P}_1)$ and $W(\mathcal{P}_2)$ respectively, then there are also adiabatic processes $\mathcal{P}$ with any value of $W(\mathcal{P})$, for $W(\mathcal{P}_1) \leq W(\mathcal{P}) \leq W(\mathcal{P}_2)$.

In order to formulate the fourth demand Carathéodory considers a special kind of adiabatic process. He argues that one can perform an adiabatic process from any given initial state to any given final shape, in such a way that the deformation coordinates follows some prescribed continuous functions of time:

$$x_1(t), \ldots, x_{n-1}(t), \tag{13}$$

Note that the system will generally not remain in equilibrium in such a process, and therefore the behaviour of the thermal coordinate x_0 remains unspecified.

Consider a series of such adiabatic processes in which the velocity of the deformation becomes 'infinitely slow', i.e. a series in which the derivatives $\dot{x}_1(t), \ldots, \dot{x}_{n-1}(t)$ converge uniformly towards zero. He calls this limit a *quasi-static change of state*. the final demand is now:

4. In a quasi-static change of state, the work done on the system converges to a value W, depending only on the given initial state and final shape, which can be expressed as a path integral along the locus of the functions (13):

$$W = \int đW = \int p_1 dx_1 + \cdots + p_n dx_{n-1},$$

where $p_1, \ldots, p_n$ denote some functions on Γ. Physically, this demand says that for adiabatic processes, in the quasi-static limit, there is no internal friction or hysteresis.

Carathéodory's version of the first law (which I have not discussed here), can then be invoked to show that

$$W = U(s_f) - U(s_i), \tag{14}$$

or in other words, the work done in quasi-static limit equals the energy difference between final and initial state. He argues that by this additional condition the value of the thermal coordinate of the final state is also uniquely fixed. Since the choice of a final shape is arbitrary, this holds also for all intermediate stages of the process. Thus, a quasi-static adiabatic change of state corresponds to a unique curve in Γ.

With this concept of a 'simple system' Carathéodory obtains the following

> CARATHÉODORY'S THEOREM: For simple systems, Carathéodory's principle is equivalent to the proposition that the differential form $đQ := dU - đW$ possesses an integrable divisor, i.e. there exist functions S and T on the state space Γ such that
>
> $$đQ = TdS. \tag{15}$$

Thus, for simple systems, every equilibrium state can be assigned a value for entropy and absolute temperature. Curves representing quasi-static adiabatic changes of state are characterised by the differential equation $đQ = 0$, and by virtue of (15) one can conclude that (if $T \neq 0$) entropy remains constant. Obviously the functions S and T are not uniquely determined by the relation (15). Carathéodory discusses further conditions to determine the choice of T and S up to a constant of proportionality, and extends the discussion to composite simple systems. However, I will not discuss this issue.

Before we proceed to the discussion of the relation of this formulation with the arrow of time, I want to mention a number of strong and weak points of the approach. A major advantage of Carathéodory's approach is that it provides a suitable mathematical formalism for the theory, and brings it in line with other modern physical theories. The way this is done is comparable to the (contemporary) development of relativity theory. There, Einstein's original approach, which starts from empirical principles like the light postulate and the relativity principle, was replaced by an abstract geometrical structure, Minkowski spacetime, where these empirical principles are incorporated in local properties of the metric. Similarly, Carathéodory constructs an abstract state space where an empirical statement of the second law is converted into a local topological property. Furthermore, all coordinate systems are treated on the same footing (as long as there is only one thermal coordinate, and they generate the same topology).[9] Note further that the environment of the system is never mentioned explicitly in his treatment of the theory. This too is a big conceptual advantage.

[9]Indeed, the analogy with relativity theory can be stretched even further. Lieb and Yngvason call the set $\mathcal{F}_s = \{t : s \prec t\}$ the 'forward cone' of s. This is analogous to the future light cone of a point p in Minkowski spacetime. (i.e. the set of all points q which are 'causally accessible' from p.) Thus, Carathéodory's principle implies that s is always on the boundary of its own forward cone.

But Carathéodory's work has also provoked objections, in particular because of its high abstraction. Many complain that the absence of an explicit reference to a *perpetuum mobile* obscures the physical content of the second law. The question has been raised (e.g. by Planck [19]) whether the principle of Carathéodory has any empirical content at all. However, Landsberg [16] has shown that for simple systems Kelvin's principle implies Carathéodory's principle, so that any violation of the latter would also be a violation of the former.

Other problems in Carathéodory's approach concern the additional assumptions needed to obtain the result (15). In the first place, we have seen that the result is restricted to simple systems, involving four additional auxiliary conditions. Falk and Jung [10] objected that the division of five assumptions into four pertaining to simple systems and one 'Principle', expressing a general law of nature, seems ad hoc. Indeed, the question whether Carathéodory's principle can claim empirical support for non-simple systems still seems to be open.

Secondly, Bernstein [1] has pointed out defects in the proof of Carathéodory's theorem. What his proof actually establishes is merely the *local* existence of functions S and T obeying (15). But this does not mean there exists a single pair of functions, defined globally on Γ, that obey (15). In fact, a purely local proposition like Carathéodory's principle is too weak to guarantee the existence of a global entropy function.

For the purpose of this essay, of course, we need to investigate whether and how this work relates to the arrow of time. We have seen that Carathéodory, like Gibbs, conceives of thermodynamics as a theory of equilibrium states, rather than processes. But his concept of 'adiabatic accessibility' does refer to processes between equilibrium states. The connection with the arrow of time is therefore more subtle than in the case of Gibbs.

In order to judge the time-reversal invariance of the theory of Carathéodory according to the criterion on page 2 it is necessary to specify a time reversal transformation R. It seems natural to choose this in such a way that $Rs = s$ and $R(\prec) = \succ$. (Since the time reversal of an adiabatical process from s to t is an adiabatic process from t to s) Then Carathéodory's principle is *not* TRI. Indeed, the principle forbids that Γ contains a 'minimal state' (i.e. a state s from which one can reach all states in some neighborhood of s. It allows models where a 'maximal' state exists, (i.e. a state s from which one can reach no other state some neighborhood. Time reversal of such a model violates Carathéodory's principle. However, this non-invariance manifests itself only in rather pathological cases. (For a fluid, a 'maximal' state would be one for which temperature and volume cannot be increased.) If we exclude the existence of such maxima, Carathéodory's theory becomes TRI.

Carathéodory also discusses the notorious notion of irreversibility. Consider, for a simple system, the class of all final states s' with a given shape $(x'_1, \ldots, x'_{n-1})$ that are adiabatically accessible from a given initial state $s = (x_0, \ldots x_{n-1})$. For example, an adiabatically isolated gas is expanded from some initial state (T, V) to

some desired final volume V'. The expansion may take place by moving a piston, slowly or more or less suddenly. The set of final states that can be reached in this fashion differ only in the value of their thermal coordinate x_0'. Due to demand 3 above, the class of accessible final states constitute a connected curve. Carathéodory argues that, for reasons of continuity, the values of S attained on this curve will also constitute a connected interval. Now among the states of the curve there is the final state, say t, of a quasi-static adiabatic change of state starting from s. And we know that $S(t) = S(s)$. Carathéodory argues that this entropy value $S(s)$ cannot be an internal point of this interval. Indeed, if it were an internal point, then there would exist a small interval $(S(s) - \epsilon, S(s) + \epsilon)$ such that the corresponding states on the curve would all be accessible from s. Moreover, it is always assumed that we can change the deformation coordinates in an arbitrary fashion by means of adiabatic state changes. By quasi-static adiabatic changes of state we can even do this with constant entropy. But then, all states in a neighborhood of s would be adiabatically accessible, which violates Carathéodory's principle.

Therefore, all final states with the final shape $(x_1', \ldots, x_{n-1}')$ that can be reached from the given point s must have an entropy in an interval of which $S(s)$ is a boundary point. Or in other words, they all lie one and the same side of the hypersurface $S = \text{const}$. By reasons of continuity he argues that this must be the same side for all initial states. Whether this is the side where entropy is higher, or lower than that of the initial state remains an open question. According to Carathéodory, a further appeal to empirical experience is necessary to decide this issue.

He concludes:

> '[It] follows from our conclusions that, when for any change of state the value of the entropy has not remained constant, one can find no adiabatic change of state, which is capable of returning the considered system from its final state back to its initial state. *Every change of state, for which the entropy varies is "irreversible".*' (Carathéodory 1909, p. 378).

Without doubt, this conclusion sounds pleasing in the ears of anyone who believes that irreversibility is the genuine trademark of the second law. But a few remarks are in order.

First, Carathéodory's conclusion is neutral with respect to time reversal: both increase and decrease of entropy is irreversible! Planck objected that the approach is not strong enough to characterise the direction of irreversible processes. In fact Carathéodory admitted this point [5]. He stressed that an additional appeal to experience is necessary to conclude that changes of entropy in adiabatic processes are always positive (if $T > 0$). In other words, in Carathéodory's approach this is not a consequence of the second law.

A second remark is that 'irreversible' here means that the change of state cannot be undone in an *adiabatic* process. This is yet another meaning for the term,

different from those we have discussed before. The question is then of course whether changes of states that cannot be undone by an adiabatic process, might perhaps be undone by some other process. Indeed, it is not hard to find examples of this possibility: consider a cylinder of ideal gas in thermal contact with a heat reservoir. When the piston is pulled out quasi-statically[10], the gas does work, while it takes in heat from the reservoir. Its entropy increases, and the process would thus qualify as irreversible in Carathéodory's sense. But Planck's book discusses this case as an example of a reversible process! Indeed, when the gas is quasi-statically recompressed, the heat is restored to the reservoir and the initial state is recovered for both system and reservoir. Thus, Carathéodory's concept of 'irreversibility' does not coincide with Planck's.

There is also another way to investigate whether Carathéodory's approach captures the content of the second law á la Clausius, Kelvin or Planck, namely by asking whether the approach of Carathéodory allows models in which these formulations of the second law are invalid. An example is obtained by applying the formalism to a fluid while swapping the meaning of terms in each of the three pairs 'heat /work', 'thermal/deformation coordinate' and 'adiabatic'/'without any exchange of work'. The validity of Carathéodory's formalism is invariant under this operation, and a fluid remains a simple system. Indeed, we obtain, as a direct analog of (15): $đW = pdV$ for all quasi-static processes of a fluid. This shows that, in the present interpretation, pressure and volume play the role of temperature and entropy respectively. Furthermore, irreversibility makes sense here too. For fluids with positive pressure, one can increase the volume of a fluid without doing work, but one cannot decrease volume without doing work. But still, the analog of the principles of Clausius of Kelvin are false: A fluid with low pressure can very well do positive work on another fluid with high pressure by means of a lever or hydraulic mechanism. And, thus, the sum of all volumes of a composite system can very well decrease, even when no external work is provided.

7 Lieb and Yngvason

Lieb and Yngvason [17] have recently provided a major contribution, by elaborate rigorous approach to the second law. In this context, I cannot do justice to their work, and will only sketch the main ideas, as far as they are relevant to my topic.

On the formal level, this work builds upon the approach of [4] and [12]. (In its physical interpretation, however, it is more closely related to Planck, as we will see below.) A system is represented by a state space Γ on which a relation $\prec$ of adiabatic accessibility is defined. All axioms mentioned below are concerned

[10]Carathéodory's precise definition of the term 'quasi-static' is, of course applicable to adiabatic processes only. I use the term here in the more loose sense of Section 2.

with this relation. Further, Lieb and Yngvason introduce a formal operation of combining two systems in state s and t into a composite system in state (s,t), and the operation of 'scaling', i.e. the construction of a copy in which all its extensive quantities are multiplied by a positive factor α. This is denoted by a multiplication of the state with α. These scaled states αs belong to a scaled state space $\Gamma_{(\alpha)}$. The main axioms of Lieb and Yngvason apply to all states $s \in \cup_\alpha \Gamma_{(\alpha)}$ (and compositions of such states). They read:

A1. REFLEXIVITY: $s \prec s$

A2. TRANSITIVITY: $s \prec t$ and $t \prec r$ imply $s \prec r$

A3. CONSISTENCY: $s \prec s'$ and $t \prec t'$ implies $(s,t) \prec (s',t')$

A4. SCALE INVARIANCE: If $s \prec t$ then $\alpha s \prec \alpha t$ for all $\alpha > 0$

A5. SPLITTING AND RECOMBINATION: For all $0 < \alpha < 1 : s \prec (\alpha s, (1-\alpha)s)$ and $(\alpha s, (1-\alpha)s) \prec s$

A6. STABILITY: If there are states t_0 and t_1 such that $(s, \epsilon t_0) \prec (r, \epsilon t_1)$ holds for a sequence of ϵ's converging to zero, then $s \prec r$.

7. COMPARABILITY HYPOTHESIS: For all states s, t in the same space Γ: $s \prec t$ or $t \prec s$. [11]

The comparability hypothesis has, as its name already indicates, a lower status than the axioms. It is intended as a characterisation of a particular type of thermodynamical systems, namely, of 'simple' systems and systems composed of such 'simple' systems.[12] A substantial part of their paper is devoted to an attempt to derive this hypothesis from further axioms. I will, however, not go into this.

The aim of the work is to derive the following result, which Lieb and Yngvason call

> THE ENTROPY PRINCIPLE (LIEB AND YNGVASON VERSION): There exists a function[13] S defined on all states of all systems such that when s and t are comparable then
>
> $$s \prec t \text{ iff } S(s) \leq S(t). \tag{16}$$

[11] The clause 'in the same space Γ' means that the hypothesis is not intended for the comparison of states of scaled systems. Thus, it is not demanded that we can either adiabatically transform a state of 1 mole of oxygen into one of 2 moles of oxygen or conversely.

[12] Beware that the present meaning of the term does not coincide with that of Carathéodory. For simple systems in Carathéodory's sense the comparability hypothesis need not hold.

[13] Actually, the Lieb-Yngvason entropy principle also states the additivity and extensivity of the entropy function.

The authors interpret the result (16) as an expression of the second law : 'It says that entropy must increase in an irreversible process.' and: 'the physical content of [(16)] . . . [is that]. . . adiabatic processes not only increase entropy but an increase in entropy also dictates which adiabatic processes are possible (between comparable states, of course).' [17, p. 19,20]).

The question whether this result actually follows from their assumptions is somewhat involved. They show that the entropy principle follows from axioms A1–A6 and the comparability hypothesis under some special conditions which, physically speaking, exclude mixing and chemical reactions. To extend the result, an additional ten axioms are needed (three of which serve to derive the comparability hypothesis). And even then, only a weak form of the above entropy principle is actually obtained, where 'iff' in (16) is replaced by 'implies'.

Before considering the interpretation of this result more closely, a few general remarks are in order. This approach combines mathematical precision, clear and plausible axioms and achieves a powerful theorem. This is true progress in the formulation of the second law. Note that the theorem is obtained without appealing to anything remotely resembling Carathéodory's principle. This is undoubtedly an advantage for those who judge that principle too abstract. In fact the axioms and hypothesis mentioned above allow models which violate the principle of Carathéodory [17, p. 91]. For example, it may be that all states are mutually accessible, in which case the entropy function S is simply a constant on Γ.

For the purpose of this paper, the question is whether there is a connection with the arrow of time in this formulation of the second law. As before, there are two aspects to this question: irreversibility and time-reversal (in)variance. We have seen that Lieb and Yngvason interpret the relation (16) as saying that entropy must increase in irreversible processes. At first sight, this interpretation is curious. Adiabatic accessibility is not the same thing as irreversibility. So how can the above axioms have implications for irreversible processes?

This puzzle is resolved when we consider the physical interpretation which Lieb and Yngvason propose for the relation $\prec$:

> ADIABATIC ACCESSIBILITY: A state t is adiabatically accessible from a state s, in symbols $s \prec t$, if it is possible to change the state from s to t by means of an interaction with some device (which may consist of mechanical and electric parts as well as auxiliary thermodynamic systems) and a weight, in such a way that the auxiliary system returns to its initial state at the end of the process whereas the weight may have changed its position in a gravitational field' [17, p. 17].

This view is rather different from Carathéodory's, or indeed, anybody else's: clearly, this term is not intended to refer to processes occurring in a thermos flask. As the authors explicitly emphasise, even processes in which the system is *heated* are adiabatic, in the present sense, when this heat is generated by an electrical

current from a dynamo driven by descending weight. Actually, the condition that the auxiliary systems return to their initial state in the present concept is strongly reminiscent of Planck's concept of 'reversible'!

This is not to say, of course, that they are identical. As we have seen before, a process $\mathcal{P}$ involving a system, an environment and a weight at height h, which produces the transition $\langle s, Z, h\rangle \xrightarrow{\mathcal{P}} \langle s', Z', h'\rangle$ is reversible for Planck iff there exists a 'recovery' process $\mathcal{P}'$ which produces $\langle s', Z', h'\rangle \xrightarrow{\mathcal{P}'} \langle s, Z, h\rangle$. Here, the states Z and Z' may differ from each other. For Lieb and Yngvason, a process $\langle s, Z, h\rangle \xrightarrow{\mathcal{P}} \langle s', Z', h'\rangle$ is adiabatic iff $Z = Z'$. But in all his discussions, Planck always restricted himself to such reversible processes 'which leave no changes in other bodies', i.e. obeying the additional requirement $Z = Z'$. These processes are adiabatic in the present sense.

A crucial consequence of this is that, in the present sense, it follows that if a process $\mathcal{P}$ as considered above is adiabatic, any recovery process $\mathcal{P}'$ is automatically adiabatic too. Thus, we can now conclude that if an adiabatic process is accompanied by an entropy increase, it cannot be undone, i.e., it is irreversible in Planck's sense. This explains why the result (16) is seen as a formulation of a principle of entropy increase In fact, we can reason as follows: assume s and t are states which are mutually comparable, and that $S(s) < S(t)$. According to (16), we then have $s \prec t$ and $t \not\prec s$. This means that there exists a process from s to t which proceeds without producing any change in auxiliary systems except, possibly, a displacement of a single weight. At the same time there exists no such process from t to s. The first-mentioned process is therefore irreversible in Planck's sense. Thus we have at last achieved a conclusion implying the existence of irrecoverable processes by means of a satisfactory argument!

However, it must be noted that this conclusion is obtained only for systems obeying the comparability hypothesis and under the exclusion of mixing and chemical processes. The weak version of the entropy principle, which is derived when we drop the latter restriction, does not justify this conclusion. Moreover, note that it would be incorrect to construe (16) as a characterisation of *processes*. The relation $\prec$ is interpreted in terms of the *possibility* of processes. As remarked in section 6, one and the same change of state can very well be obtained (or undone) by means of different processes, some of which are adiabatic and others not. Thus, when $S(s) < S(t)$ for comparable states, this does not mean that *all* processes from s to t are irreversible, but only that there exists an adiabatic irreversible process between these states. So the entropy principle here is not the universal proposition of Planck.

The next question concerns the time-reversal (in)variance of this approach. As before, we can look upon the axioms as singling out a class of possible worlds $\mathcal{W}$. It is easy to show, using the implementation of time reversal used earlier, i.e. replacing $\prec$ by $\succ$, the six general axioms, and the comparability hypothesis, are

TRI![14] The fact that it is not necessary to introduce time-reversal non-invariance into the formalism to obtain the second law, is very remarkable.

However, there remains one problematical aspect of the proposed physical interpretation. It refers to the state of auxiliary systems in the environment of the system. Thus, we are again confronted by the old and ugly question, when shall we say that the state of such auxiliary systems has changed, and when are we fully satisfied that their initial state is restored. This question remains rather intractable from the point of view of thermodynamics, when one allows arbitrary auxiliary systems (e.g. living beings) whose states are not represented by the thermodynamical formalism. Thus, the question when the relation $\prec$ holds cannot be decided in thermodynamical terms.

8 DISCUSSION

We have seen that there is a large variety in the connections between irreversibility and the second law. On one end of the spectrum, there is Planck's view that the second law expresses the irreversibility of all processes in Nature. A convincing derivation of this bold claim has, however, never been given. On the other extreme, we find Gibbs' approach, which completely avoids any connection with time.

But even for approaches belonging to the middle ground, the term 'irreversible' is used in various meanings, time-reversal non-invariant, irrecoverable, and quasi-static. In the long-standing debate on the question how the second law relates to statistical mechanics, however, most authors have taken irreversibility in the sense of time-reversal non-invariance. The point that in thermodynamics, the term usually means something very different has been almost completely overlooked.

The more careful and formal approaches by Carathéodory and, in particular Lieb and Yngvason rather yield a surprising conclusion. It is possible to build up a precise formulation of the second law without introducing a non-TRI element in the discussion. The resulting formalism, therefore, remains strictly neutral to the question of whether entropy increases or decreases. It implies only that an entropy function can be constructed consistently, i.e. as either increasing between adiabatically accessible states of *all* simple systems, or decreasing. At the same time, the Lieb-Yngvason approach does imply that entropy increasing processes between comparable states are irreversible in Planck's sense. This, of course, shows once more the independence of the two notions.

[14] This conclusion cannot be extended to the complete set of axioms proposed by Lieb and Yngvason. In particular, their axioms A7 and T1, which address mixing and equilibration processes, are explicitly non-TRI. (I thank Jakob Yngvason for pointing this out to me.) However, these axioms are needed only in the derivation of the (TRI) comparability hypothesis, and not in the derivation of the entropy principle.

Finally, I would like to point out an analogy between the axiomatisation of thermodynamics in the Carathéodory and Lieb-Yngvason approach and that of special relativity in the approach of Robb [20]. In both cases, we start out with a particular relationship $\prec$ which is assumed to exist between points of a certain space. In relativity, this is this is the relation of connectability by a causal signal. In both cases, it is postulated that this relation forms a pre-order. In both cases, important partial results show that the forward sectors $C_s = \{t \ : s \prec t\}$ are convex and nested and that s is on the boundary of C_s. And in both cases the aim is to show that the space is 'orientable' [9] and admits a global function which increases in the forward sector. If this analogy is taken seriously, the Lieb-Yngvason entropy principle has just as much to do with TRI as the fact that Minkowski space-time admits a global time coordinate.

There is, however, also an important disanalogy. In thermodynamics, the space Γ represents the states of a system, and an important feature is that we can combine systems into a composite, or divide one into subsystems. In relativity, the space represent the whole of space-time, and there is no question of combining several of these.

Due to the possibility of combining systems, the Lieb-Yngvason entropy principle does yield an additional result: entropy can be defined consistently, in the sense just mentioned. For some, this result may be sufficient to conclude that the principle does express some form of irreversibility. However, as has been emphasized by Schrödinger [21] a formulation of the second law which states that all systems can only change their entropy in the same sense, is not in contradiction to the time-reversal invariance of an underlying microscopic theory.

REFERENCES

[1] B. Bernstein, Proof of Carathéodory's local theorem and its global application to thermostatics, *J. Math. Phys.*, 1:222–224, 1960.

[2] H.A. Buchdahl, *The Concepts of Classical Thermodynamics*, Cambridge University Press, Cambridge, 1966.

[3] H.B. Callen, *Thermodynamics*, Wiley, New York, 1960.

[4] C. Carathéodory, Untersuchungen über die Grundlagen der Thermodynamik, *Math. Ann.*, 67:355–386, 1909.

[5] C. Carathéodory, Über die Bestimmung der Energie und der absoluten Temperatur mit Hilfe von reversiblen Prozessen, *Sitzungsber. der Preuss. Akad. Wiss.*, 39–47, 1925.

[6] R. Clausius, *Abhandlungungen über die mechanische Wärmetheorie* Vol. 2, Vieweg, Braunschweig, 1867.

[7] R. Clausius, *Die mechanische Wärmetheorie*, Vieweg, Braunschweig, 1876.

[8] K. Denbigh, The many faces of irreversibility, *Brit. J. Phil. Sci.*, 40:501–518, 1989.

[9] J. Earman, An attempt to add a little direction to "the problem of the direction of time" *Phil. Sci.*, 41:15–47, 1974.

[10] G. Falk and H. Jung, Axiomatik der Thermodynamik, In S. Flügge, ed., *Handbuch der Physik*, Vol. III/2. Springer, Berlin, 1959.

[11] J.W. Gibbs, *The Scientific Papers of J. Willard Gibbs, Vol. 1, Thermodynamics*, Longmans, London, 1906.

[12] R. Giles, *Mathematical Foundations of Thermodynamics*, Pergamon, Oxford 1964.

[13] H.B. Hollinger and M.J. Zenzen, *The Nature of Irreversibility*, Reidel, Dordrecht, 1985.

[14] R. Illner and H. Neunzert The concept of irreversibility in the kinetic theory of gases, *Transp. Th. Stat. Phys.* 16:89-112, (1987).

[15] J. Kestin, *The Second Law of Thermodynamics*, Dowden, Hutchinson and Ross, Stroudsburg, Penn., 1976.

[16] P.T. Landsberg, A deduction of Carathéodory's principle from Kelvin's principle, *Nature*, 201:485–486, 1964.

[17] E.H Lieb and J. Yngvason, The phyics and mathematics of the second law of thermodynamics, *Phys. Rep.*, 310:1–96, 1999. Erratum, **314** (1999) 669. Also http://xxx.lanl.gov/abs/cond-mat/9708200.

[18] M. Planck, *Vorlesungen über Thermodynamik*, Veit, Leipzig, 1897.

[19] M. Planck, Über die Begrundung des zweiten Hauptsatzes der Thermodynamik, *Sitzungsber. Preuss. Ak. Wiss.*, 453–463, 1926.

[20] A.A. Robb, *The Absolute Relations of Time and Space*, Cambridge university Press, Cambridge, 1921.

[21] E. Schrödinger, Irreversibility, *Proc. Roy. Irish Ac.*, 53 A:189–195, 1950.

[22] C. Truesdell, What did Gibbs and Carathéodory leave us about thermodynamics? In J. Serrin, ed., *New Perspectives in Thermodynamics*, 101–123. Springer, Berlin, 1986.

[23] J.Uffink, Bluff your way in the second law of thermodynamics *Stud. Hist. Phil. Mod. Phys.*, to appear.

[24] J.D. van der Waals and Ph. Kohnstamm, *Lehrbuch der Thermostatik*, Barth, Leipzig, 1927.