

# Irreversibility and the second law

Jos Uffink

*Institute for History and Foundations of Science, PO Box 80.000 3508 TA Utrecht, the Netherlands*

**Abstract.** The relation between the second law of thermodynamics and the notion of irreversibility is analysed by distinguishing three different meanings of the latter and studying how they figure in several versions of the second law. A more extensive discussion is given in [1].

## 1. THREE CONCEPTS OF (IR)REVERSIBILITY

Many physical theories employ a state space  $\Gamma$  containing all possible states  $s$  of a system. A process is represented as a parameterised curve:

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

Usually a theory allows only a subclass, say  $\mathcal{W}$ , of such processes (e.g. the solutions of the equations of motion). Let  $R$  be an involution (i.e.  $R^2s = s$ ) that turns 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, where the state does not contain velocity-like parameters, one may take  $R$  to be the identity. Further, define the time reversal  $\mathcal{P}^*$  of process  $\mathcal{P}$  by:

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

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

$$\mathcal{P} \in \mathcal{W} \implies \mathcal{P}^* \in \mathcal{W}. \quad (1)$$

According to this definition the form of the laws (and a choice for  $R$ ) determines whether a theory is TRI or not. Note that it is irrelevant here whether the processes  $\mathcal{P}^*$  actually occur, but only that the theory allows them. Thus, the fact that the sun never rises in the west doesn’t mean that celestial mechanics is non-TRI.

Is time-reversal (non)invariance related to the second law? Applying the criterion to thermodynamics is no matter of routine. In contrast to mechanics, thermodynamics does not have equations of motion. Indeed, thermodynamical processes typically occur after external intervention on the system (e.g.: removing a partition, pushing a piston, etc.) and do not reflect its autonomous behaviour. Yet, classical thermodynamics, in the formulation of Clausius, Kelvin or Planck, is concerned with processes, and its second law is clearly not TRI. But in other formulations this is less clear.

Now, ‘(ir)reversible’ is an attribute of processes, not theories or laws. But in philosophy of physics, it is closely connected with TRI. Indeed, 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  $\mathcal{P}$  exists only if the theory in question is not TRI. Conversely, every non-TRI theory admits irreversible processes in this sense. Therefore, discussions about (ir)reversibility and (non)-TRI in philosophy of physics mostly coincide.

However, the thermodynamics literature often uses the term ‘irreversibility’ to denote processes one might also call *irrecoverable*, i.e., when 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. Wear and tear, erosion etc. are the obvious examples. This is the sense of irreversibility that Planck intended, when he called it the essence of the second law.

(Ir)recoverability differs from (non)-TRI in at least two respects. First, the only thing that matters here is the retrieval of the initial state  $s_i$ . It is not necessary to find a process  $\mathcal{P}^*$  retracing the intermediate stages of the original process in reverse order. Secondly, we are dealing with a *complete* recovery. This means that all auxiliary systems that may have been used in the original process are also returned to their initial state.

A schematic expression of the idea is this. 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 transition:

$$\langle s_i, Z_i \rangle \xrightarrow{\mathcal{P}} \langle s_f, Z_f \rangle \quad (2)$$

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

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

The term ‘reversible’ is also used in a third sense, to denote processes which proceed so slowly that the system remains in equilibrium ‘up to a negligible error’ during the entire process. This is the meaning embraced by Clausius, and it appears to be the most common usage of the term in the physical-chemical literature; see e.g. [2, 3]. A more apt name for this kind of processes is *quasi-static*. Of course, the above characterisation is vague, and has to be amended by specifying what ‘errors’ are meant and when they are ‘small’. These criteria invoke a limit procedure so that, strictly speaking, reversibility is here not an attribute of one process but of a series of processes.

Quasi-static processes need not be the same as those called reversible in the previous two senses. An ideal harmonic oscillator is reversible in Planck’s sense, but not quasi-static. Conversely, the discharge of a condenser through a large resistance can be made to proceed quasi-statically, but irreversibly 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. Consider a process  $\mathcal{P}_N$  in which a system, originally at temperature  $\theta_1$  is consecutively put 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  $\theta_2$ . 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  $\mathcal{P}_N$  is impossible.

Nevertheless, many authors call such a curve ‘reversible’, because one may consider other processes  $\mathcal{Q}_N$ , in which the system, originally at temperature  $\theta_2$ , is put in contact with a series of heat baths, each slightly *colder* than the previous. Again, each  $\mathcal{Q}_N$  is

non-TRI. *A fortiori*, no  $\mathcal{Q}_N$  is the time-reversal of  $\mathcal{P}_N$ . Yet, if we take the quasi-static limit, the state change traverses the same curve in equilibrium space as in the previous case, in opposite direction. The point is, of course, that precisely because this curve is not itself a process, the notion of time reversal does not apply to it.

## 2. PLANCK

Planck has always argued that the second law expresses irrecoverability of all processes in nature. However, it is not easy to analyse his arguments for this claim. Various editions of his book [4] differ in many decisive details. Also, the English translation is unreliable. It uses the term ‘reversible’ indiscriminately, where Planck distinguishes between *umkehrbar*, which he uses in Clausius’ sense, i.e. meaning ‘quasi-static’, and *reversibel*, in the sense of Kelvin (1852) meaning ‘recoverable’. Moreover, he presented a completely different argument from the eighth edition onwards.

I shall only mention Planck’s latter argument, published first in [5]. He starts from the statement that “friction is an *irreversibel* process”, which he considers to be an expression of Kelvin’s principle.<sup>1</sup> He then considers an adiabatically isolated fluid 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

$$(s, Z, h) \longrightarrow (s', Z, h').$$

He argues that, by means of ‘*reversibel-adiabatic*’<sup>2</sup> 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

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

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  $U(s^*) = U(s')$ . Since  $U$  and  $V$  determine the state of the fluid completely,  $s^*$  and  $s'$  must coincide.
- (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

---

<sup>1</sup> 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 cools a reservoir and does work by means of any type of auxiliary system that operates in a cycle.

<sup>2</sup> Apparently, Planck’s pen slipped here. He means: *umkehrbar-adiabatic*.

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, 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

$$dU = TdS. \quad (4)$$

Assuming that  $T > 0$ , it follows that  $U$  must vary in the same sense as  $S$ . That is, the three cases 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 argument) 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 [5, p. 463]

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  $dQ = TdS$ ; and of a rather generous supply of quasi-static processes.

### 3. CARATHÉODORY

Carathéodory [6] construed thermodynamics as a theory of equilibrium states rather than (cyclic) processes. A thermodynamical system is described by a state space  $\Gamma$ , represented as a (subset of a)  $n$ -dimensional manifold with the state variables serving as coordinates. He assumes that  $\Gamma$  is equipped with the standard Euclidean topology. But metrical properties do not play a role, and there is no preference for a particular system of coordinates. The fundamental concept is a relation called *adiabatic accessibility*, which represents whether state  $t$  can be reached from state  $s$  in an adiabatic process,<sup>3</sup> and the second law is formulated as follows:

CARATHÉODORY'S PRINCIPLE: In every open neighborhood  $U_s \subset \Gamma$  of every state  $s$  there are states  $t$  such that for some open neighborhood  $U_t$  of  $t$ : all states  $r \in U_t$

---

<sup>3</sup> Carathéodory calls a process adiabatic if it takes place in a container such that the system 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 an adiabatic process is by deformation of the walls. (E.g. compression or stirring.)

cannot be reached adiabatically from  $s$ .

He then introduces so-called ‘simple systems’ (whose definition I skip) and obtains

CARATHÉODORY’S THEOREM: For simple systems, Carathéodory’s principle is equivalent to the proposition that the differential form  $dQ := dU - dW$  possesses an integrable divisor, i.e. there exist functions  $S$  and  $T$  on  $\Gamma$  such that  $dQ = TdS$ .

Thus, for simple systems, each equilibrium state has an entropy and absolute temperature. Curves representing quasi-static adiabatic changes of state are characterised by the differential equation  $dQ = 0$ , and (if  $T \neq 0$ ) their entropy remains constant.

Let me mention a few strong and weak points of the approach. An advantage is that it provides exact formalism for thermodynamics, comparable to relativity theory. There, Einstein’s approach, starting from empirical principles (the light postulate and relativity principle), was replaced by an abstract Minkowski spacetime, where those principles are incorporated in local properties of the metric. Similarly, Carathéodory constructs an abstract state space which converts an empirical statement of the second law into a local topological property. Further, all coordinate systems are on the same footing (as long as there is only one thermal coordinate, and they generate the same topology).

But Carathéodory’s work has also provoked objections. Many complain that the lack of explicit reference to a *perpetuum mobile* obscures the physical content of the second law. Other problems in Carathéodory’s approach concern the additional assumptions, i.e. the restriction to simple systems, used to obtain the theorem. Further, Carathéodory’s proof merely establishes the *local* existence of functions  $S$  and  $T$ . It does not guarantee the existence of a pair of globally defined functions that obey  $dQ = TdS$ .

For the present purpose, the question is whether and how this work relates to ‘irreversibility’. Carathéodory 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.

In order to judge the time-reversal invariance of the theory of Carathéodory one must specify a time reversal transformation  $R$ . It seems natural to choose this in such a way that  $Rs = s$  and  $R(\prec) = \succ$ . Then Carathéodory’s principle is *not* TRI. Indeed, the principle forbids that  $\Gamma$  contains a state  $s$  from which one can reach all states in some neighborhood of  $s$ . It allows models where a state  $s$  exists from which one can reach no other state in some neighborhood. Time reversal of such a model violates Carathéodory’s principle. However, this non-invariance manifests itself only in rather pathological cases. If we exclude them, Carathéodory’s theory becomes TRI.

It is easy to show that Carathéodory’s approach fails to capture the content of the second law à la Planck, namely by exhibiting models of his formalism in which this version of the second law is invalid. An example is obtained by 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 for fluids. Indeed, we obtain analogously,  $dW = pdV$  for all quasi-static processes of a fluid. Thus pressure and volume here play the role of temperature and entropy respectively. Further, 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.

#### 4. LIEB AND YNGVASON

Lieb and Yngvason [7] recently provided a rigorous approach to the second law. I can only sketch those main ideas that are relevant to my topic. Formally, their work builds upon the approach of [6] and [8]. (In its physical interpretation, however, it is more closely related to Planck.) A system is represented by a state space  $\Gamma$  on which a relation  $\prec$  of adiabatic accessibility is defined. Further, one may combine two systems in state  $s$  and  $t$  into a composite system in state  $(s, t)$ , and there is an operation of ‘scaling’, i.e. the construction of a copy in which all its extensive quantities are increased by a factor  $\alpha$ . This is denoted as multiplying the state by  $\alpha$ . The main axioms 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  $0 < \alpha < 1$ :  $s \prec (\alpha s, (1 - \alpha)s) \prec s$
  - 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  $\epsilon$ 's converging to zero, then  $s \prec r$ .
7. COMPARABILITY HYPOTHESIS: For all states  $s, t$  in the same  $\Gamma$ :  $s \prec t$  or  $t \prec s$ .<sup>4</sup>

The comparability hypothesis intends to characterise a particular type of ‘simple’ systems.<sup>5</sup> A major part of their paper is devoted to to derive this from further axioms.

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

THE ENTROPY PRINCIPLE: There exists a function  $S$  defined on all states of all systems such that when  $s$  and  $t$  are comparable then

$$s \prec t \text{ iff } S(s) \leq S(t). \quad (5)$$

It is shown that this follows from axioms A1–A6 and the comparability hypothesis under conditions which, physically speaking, exclude mixing and chemical reactions.

Note that the theorem is obtained without appealing to Carathéodory’s principle. In fact, the approach allows models which violate Carathéodory’s principle. For my purpose, the question is what connection is with irreversibility in this formulation of the

---

<sup>4</sup> The clause ‘in the same  $\Gamma$ ’ means that the hypothesis is not intended for the comparison of states of scaled systems. Thus, it is not demanded that we can adiabatically transform 1 mole of oxygen into 2 moles of oxygen or conversely.

<sup>5</sup> 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.

second law. As before, there are two aspects to this question: irrecoverability and time-reversal (in)variance. Lieb and Yngvason interpret the relation (5) as saying that entropy must increase in irreversible processes. At first sight, this interpretation is curious: Is adiabatic accessibility the same thing as irreversibility?

This puzzle is resolved by looking at the physical interpretation Lieb and Yngvason propose for  $\prec$ :

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' [7, p. 17].

This view differs from Carathéodory's, or indeed, anybody else's: clearly, this term is not intended to refer to processes occurring in a thermos flask. 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 reminiscent of Planck's concept of 'reversible'!

This is not to say, of course, that they are identical. A process  $\mathcal{P}$  involving a system, an environment and a weight at height  $h$ , producing 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  $\langle s', Z', h' \rangle \xrightarrow{\mathcal{P}'} \langle s, Z, h \rangle$ .

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 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 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 may 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 (5) is seen as a formulation of a principle of entropy increase. Thus we obtain the existence of irrecoverable processes by means of a satisfactory argument!

However, note that it would be incorrect to construe (5) as a characterisation of *processes*. The relation  $\prec$  is interpreted in terms of the *possibility* of processes. Thus, when  $S(s) < S(t)$  for comparable states, this does not mean that *all* processes from  $s$  to  $t$  are irrecoverable, but only that there exists an adiabatic irrecoverable process between these states. So the entropy principle here is not the same as Planck's.

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!<sup>6</sup> Quite remarkably, no time-reversal non-invariance is needed to obtain this version of the second law.

## 5. DISCUSSION

There is much variety in views on irreversibility and the second law. On one end, Planck maintained that this law expresses the irrecoverability of all processes in Nature. A demonstration of this bold claim has, however, never been given. On the other extreme is Gibbs, who completely avoided any connection with time.

But even for approaches in the middle, the term ‘reversible’ is used in various meanings: time-reversal invariance, recoverable, and quasi-static. In the debate on the relation of the second law to statistical mechanics, however, most authors have taken irreversibility in the sense of time-reversal non-invariance. The point that in thermodynamics the term might mean something different has been almost completely overlooked.

The formal approaches by Carathéodory and Lieb and Yngvason show that it is possible to build up a precise formulation of the second law without introducing a non-TRI element. The resulting formalism 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 shows once more the independence of the two notions.

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 [10]. 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 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  $\mathcal{C}_s = \{t : s \prec t\}$  are convex and nested and that  $s$  is on the boundary of  $\mathcal{C}_s$ . 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.

## REFERENCES

1. Uffink, J. *Studies in History and Philosophy of Modern Physics*, **32** 305–394 (2001).
2. Hollinger, H., and Zenzen, M., *The Nature of Irreversibility*, D. Reidel, Dordrecht, 1985.
3. Denbigh, K., *British Journal for Philosophy of Science*, **40**, 501–518 (1989).
4. Planck, M., *Vorlesungen über Thermodynamik*, Verlag von Veit & Comp., Leipzig, 1897.
5. Planck, M., *Sitzungsberichte der Preussischen Akademie der Wissenschaften*, pp. 453–463 (1926).
6. Carathéodory, C., *Mathematische Annalen*, **67**, 355–386 (1909).

---

<sup>6</sup> Some axioms used by Lieb and Yngvason are explicitly non-TRI. However, these are needed only in their derivation of the (TRI) comparability hypothesis, and not for the entropy principle.

7. Lieb, E., and Yngvason, J., *Physics Reports*, **310**, 1–96 (1999), erratum, **314** (1999) 669.
8. Giles, R., *Mathematical Foundations of Thermodynamics*, Pergamon, Oxford, 1964.
9. Earman, J., *Journal of Philosophy*, **64**, 543–549 (1967).
10. Robb, A.A. *The Absolute Relations of Time and Space* Cambridge university Press, Cambridge, 1921.