Heylighen F. (1989): "Causality as Distinction Conservation: a theory of predictability, reversibility and time order", Cybernetics and Systems 20, p. 361-384.

CAUSALITY AS DISTINCTION CONSERVATION: a theory of predictability, reversibility and time order

Francis HEYLIGHEN[1]
Transdisciplinary Research Group
Free University of Brussels (VUB)

ABSTRACT. "Equal causes have equal effects" is reformulated by defining causality as a distinction-conserving relation. Unpredictable, respectively irreversible, processes are analysed as processes in which distinctions are created, respectively are destroyed. Different types of partially causal and pseudo-causal relations are examined. Time order is derived from distinction conservation. It is argued that the emergence of macroscopic distinctions and causal relations is due to a self-organizing evolution, characterized by natural selection. The relationship between "physical" and "observer-dependent" factors in determining causal relations is discussed.

1. Introduction

Different formulations of what is the essence of scientific knowledge all seem to require the concept of a causal relation. If we define the aim of science to be prediction, then we must assume some kind of relationship which allows to infer future events (effects) from past or present events (causes). If we consider explanation to be the objective of science, then we presuppose that it is possible to uncover some mechanism by which present situations (effects) could be derived from past conditions (causes).

It is clear then that an analysis of scientific knowledge requires a basic understanding of causality. We know, however, that the definition of causality constitutes a very old philosophical problem, dating back at least as far as Aristotle. Many different characterizations of causality have been given in the course of centuries, but none seems able to encompass all the relevant features. Let me summarize here some of the basic questions, to which different models have provided different answers.

All philosophers seem to agree that causality is a *relation*, relating "causes" to "effects". In order to understand the nature of this relation we must first unambiguously determine the *relata*, the entities between which the relation exists. Here we come to a first major problem :

[1] *address*: PESP, Vrije Universiteit Brussel, Pleinlaan 2, B-1050 Brussels, Belgium,
tel. (32) (0) 2-641 25 25

what kind of entities are causes and effects? Are they objects, events, properties, processes, propositions, situations, percepts, ideas ...? Different theories of causality will in general choose different categories for the relata. For example, logical theories may consider a causal relation as a special type of implication between propositions. The theory of Hume views causality as an association between experiences.

The second main problem concerns the nature of the connection between the relata. A first essential feature is the "strength" of the determination : in how far does an effect necessarily follow from a cause, and conversely, in how far is a specific cause necessary for an effect to occur? Can the same effect result from different causes? This problem is related to that of determinism : if we assume that all events are causally related, and that causal determination is complete, then we must conclude that the world is completely determined, and that all future events are in principle predictable, given a complete knowledge of the present situation (cfr. the demon of Laplace). The second feature is that of temporal order : must a cause always precede its effect? Is synchronous or cyclic causality possible? One possible answer is given by relativity theory, which states that a causal signal can only travel forward in time, with a velocity which is less than or equal to the speed of light, but this of course invites the question : why?

A last basic problem concerns the ontological status of causality : is causality "real" or "objective", does it exist independently of the observer (ontic interpretation)? Or is it merely a cognitive construct, an association between subjective ideas or experiences (epistemic interpretation)?

These problems of interpretation are not confined to philosophy. In scientific practice many conceptual difficulties arise due to an ambiguous conception of causality. Perhaps the most famous examples of these difficulties are the so-called paradoxes of quantum mechanics. E.g. in the EPR-paradox (Einstein, Podolsky and Rosen, 1935; Jammer, 1974) we are confronted with a seeming inconsistency between two characteristics of causality : on the one hand there appears to be a direct "influence" of one event on another one, apparently implying a causal connection; on the other hand, this influence seems to travel instantaneously over a finite distance, contradicting the causality principle as it is formulated in relativity theory.

I will now propose a new approach to these classical problems. The approach is original because it introduces a new type of relata: *distinctions*. It will be shown that with such relata the causal relation becomes very simple: a distinction conserving mapping. I shall then try to formulate an answer to all the basic questions, starting from these foundations. I hope to show that such an approach allows to integrate the different aspects of causality in a quite simple and intuitive way.

2. A definition of causality

We shall start from a very simple characterization of causality :

equal causes have equal effects, (1)

and analyze it further. The first question to arise is : when are two causes or effects considered to be equal? Clearly "equal" does not mean "identical". Otherwise the principle would read (with c_1 as cause and e_1 as effect) : if $c_1 = c_1$, then $e_1 = e_1$, which is tautological. Moreover such a definition would be completely useless if you want to make predictions :

since a new cause c_2 by definition can never be identical to a previous cause c_1, the knowledge that c_1 led to e_1 does not allow us to infer anything about the effect of c_2.

One possibility might be to interpret "equal" as "similar", and to assume that if c_1 is similar to c_2, then the effect e_2 of c_2 will be similar to e_1. The problem with similarity is that it is not transitive : perhaps c_1 is similar to c_2, and c_2 similar to c_3, and c_3 to c_4, ... to c_n ; yet c_1 will in general no longer be "similar" to c_n. "Equal", on the other hand, seems to imply an equivalence relation, which is by definition transitive.

Similarity could be viewed as a measure of distance in some topological space of causes : c_i may be "close" to c_{i+1} (i=1,...,n) but for large n, c_1 will no longer be "close" to c_n. The requirement that causes which are "close" to each other be mapped upon effects which are close to each other, is really a requirement of *continuity*, and not of causality. A function mapping causes to effects may well be discontinuous, without therefore being non-causal.

A much simpler way of determining "equality" would be to consider two causes as "equal" if there is no reason to actively *distinguish* them. This is an example of *default reasoning* : if there is no indication that something would be different from the "normal" situation, then we may assume by default that knowledge about the normal case is appropriate. Models from cognitive science and artificial intelligence have shown that default reasoning is a very natural cognitive mechanism, without which an intelligent system would be unable to adapt to a complex world : it would be forced to doubt about everything, and to examine so many alternatives (most of which are highly improbable), that it would be unable to reach a decision within a finite time.

Such an assumption of cognitive economy (or "laziness", if you prefer) leads us to shift from "equality" to "distinction" as the essential, active principle for structuring our perception of causes and effects. If we assume that the causal principle (1) is valid in in both directions (i.e. we have also that "equal effects have equal causes"), then we may infer the following equivalent principle :

distinct causes have distinct effects (and vice-versa) (2)

If distinctness is represented by "≠", we may write (with e_i as the effect corresponding to cause c_i) :

$$c_1 \neq c_2 \quad \Leftrightarrow \quad e_1 \neq e_2. \tag{3}$$

Considered in this way "distinctness" is a relation. The process or connection which leads from causes to effects can now be characterized by the fact that it *conserves* this relation : the distinction between c_1 and c_2 is (bijectively) mapped upon the distinction between e_1 and and e_2. The phenomena themselves may change during the process (from c to e), but their relation of distinction remains.

Apart from the fact that it is by definition antireflexive, there is no general way to characterize a distinction relation. We will therefore direct our attention to a more restricted, simpler model of distinctness. We shall assume that all phenomena which are not distinguished belong to the same class c, and that this class can be separated from its complement or negation $c\tilde{}$, i.e. the class of all phenomena which are distinguished from those of c. The distinction between both classes can be represented by the one couple (c, $c\tilde{}$). This is much simpler than the set of all couples of mutually distinct elements, which would be needed to define a distinction relation in a set theoretic framework.

Moreover, a distinction of this type can be defined implicitly by the two axioms of

Spencer-Brown's *distinction algebra* (1969). Spencer-Brown has shown that this algebra is isomorphic to Boolean algebra, although its structure is much simpler. This shows that a distinction can also be interpreted as a Boolean variable c, with two truth values 1 (yes, the phenomenon belongs to class c) and 0 (no, it does not belong to c, but to its complement c˜). Henceforth we shall consider a distinction as a primitive entity, to be interpreted as an *element of structuration* (cognitive or physical), which is not explicitly defined, but which obeys implicitly the axioms of Boolean algebra (Heylighen, 1987, 1988). This also means that in cases where more than two alternatives must be examined, we may combine different binary distinctions by using conjunction or disjunction.

Causality can then be analysed as a mapping between such elements, to be represented in the following way :

$C : (a, a\tilde{}) \rightarrow (b, b\tilde{})$, or representing couples by single letters: (4)
$C : A \rightarrow B$

This is to be interpreted as : a causes b, a˜ (the absence of a) causes b˜ (the absence of b); the distinction A is causally related to the distinction B.

Let us now try to clarify the significance of these assumptions by studying some examples from physics.

3. Causal and non-causal processes in physics.

The physical theory of *classical mechanics* is the prototype of a completely deterministic and completely causal theory. From the state s(t) of a mechanical system at a time t, we can compute all future states s(t + T) and previous states s(t - T'). The state at a specific instant (which contains the position of the system in phase space, and the values of all the dynamical variables which determine the Hamiltonian) completely determines all past or future states. Such a state s(t) may be conceived as a cause leading to the effect s(t + T) at a later time. The determination in both directions (future and past) clearly implies a complete distinction conservation. Indeed, two states which are distinct at time t :

$s(t) \neq s'(t)$

will remain distinct at all later times t + T, and are the results of distinct states at all previous time t - T'. If the evolution of the system is represented by its possible trajectories :

$\{ s(t) \mid -\infty < t < +\infty \}$,

this means that the two trajectories s and s' will never intersect.

This is no longer true if we go from classical mechanics to *thermodynamics*. The second law of thermodynamics tells us that a closed system will evolve to the state with maximal entropy, whatever the state the system initially had. In other words, distinct initial states may lead to the same final state. This phenomenon is called *equifinality*. It leads to the *irreversibility* of the evolution : given the final state with maximal entropy, it is impossible to know which initial state led to this result, hence the initial state cannot be reconstructed, unless information is added (i.e. entropy is diminished by external input). The set of trajectories corresponding to such an evolution can be represented as a tree (upside-down) where more and more branches (initially distinct trajectories) come together to be finally merged, forming the root of the tree (the further trajectory of the maximal-entropy state).

Let us analyse this phenomenon from the viewpoint of causality. Clearly we have that

distinct causes have equal (= non-distinct) effects, contradicting the causality principle (2). Yet we still may assume that a given cause will have a given effect, i.e. the process is predictable. In fact, only one "half" of the causality principle is violated. It is still true that distinct effects (e.g. configurations with different maximal entropy states) have distinct causes. In other words the cause is sufficient to determine the effect, but it is not necessary. This may be summarized by noting that the double implication of (3), is reduced to a single implication :

$$c_1 \neq c_2 \qquad e_1 \neq e_2. \tag{5}$$

Such a relation, in which distinction may be destroyed, but not created, i.e. it is only conserved in one direction : from effects to causes (i.e. backward), corresponds to a weaker form of causality. The process is predictable, but not retrodictable (i.e. not reversible).

In thermodynamical situations which are "far from equilibrium" (Prigogine, 1979), we may encounter the opposite phenomenon : *bifurcation*. Here we have a state which is unstable in the sense that the slightest fluctuation may push the system into one out of several new regimes. Since the fluctuation is by definition unobservable, it appears to the observer that the same initial state may lead to distinct subsequent states : the trajectory of the system bifurcates. Hence the process is no longer predictable : we do not know which of the available trajectories the system will choose at the bifurcation point. The process becomes stochastic. Let us assume for simplicity that the process is still reversible (which is not very realistic in thermodynamics). In that case, we have again a violation of one half of the causality principle : equal causes may have distinct effects, yet distinct causes always have distinct effects. In other words :

$$c_1 \neq c_2 \qquad e_1 \neq e_2. \tag{6}$$

This is again a weaker form of the causality principle, complementary to the previous one, in which a cause is necessary for the corresponding effect to occur, but not sufficient. Distinctions may be created, but not destroyed, i.e. they are conserved only in the forward direction, from causes to effects.

As we remarked, the general phenomenon in far-from-equilibrium thermodynamics is neither reversible nor predictable. In this case both halves of the causality principle are violated, the process is undetermined in both directions : to the future and to the past.

This does not mean that it is impossible to predict or to explain the process in any of its aspects, however. The present definition of causality allows *partially causal* processes, i.e. processes in which some distinctions are conserved (in forward direction, in backward direction or in both), whereas others are not. For example, in general, the number of possible final states after a bifurcation will be much smaller than the total number of distinct states the system can have. An irreversible and stochastic thermodynamic system may still be characterized by energy conservation, which entails that states with distinct energies will remain distinct in this aspect. Remark also that in a stochastic or probabilistic description the distinctions between different (transition) probabilities will in general be conserved. The lack of distinction conservation at the level of states is here (partially) compensated by distinction conservation at the level of probabilities of states.

Another example of a physical theory in which absolute causality is violated, is *quantum mechanics* (cfr. Jammer, 1974). Although the dynamical evolution of quantum states, as determined by the Schrödinger equation, is perfectly deterministic, there is another

quantum mechanism which is not : the observation process. In the quantum formalism an observation process is represented by a projection operator which projects a quantum state onto an eigenstate of the observable which is measured. Since in general there are different eigenstates which are not orthogonal to the initial state to be measured, the process can have several, distinct results. We can only determine the probability with which the initial state (cause) will lead to a particular final state (effect). Clearly the process is unpredictable, i.e. equal causes can have distinct effects. But the process is also irreversible, i.e. distinct causes can have equal effects. Indeed, two distinct states on which the same observation is performed may be projected onto the same eigenstate. Hence the causality principle is violated in both directions.

A traditional illustration of this indeterminism of quantum mechanics is the process of radioactive decay. Consider two radioactive nuclei which are in all respects indistinguishable : they belong to the same type (e.g. a particular isotope of uranium), and they are in the same quantum mechanical state. Yet in general they will decay at a different time. In principle there is no way to determine which of the nuclei will decay first, or even to determine whether a given nucleus will have decayed at a certain instant of time or not. This example shows that the intuitive assumption of causality, expressed in the form (1) or (2), cannot be universally applied to physical phenomena. This invites the question : in which situations is causality valid, and in which situations is it not? Why? I hope to give some clues for answering this question in the following sections.

A quite different use of causality is made in *relativity theory*. The essence of relativity theory can be found in what is called the "causal structure" of space-time (Kronheimer & Penrose, 1967; Reichenbach, 1958). This structure is based on the "light cone" which determines a constraint on causal processes : no causal signal can travel faster than light, i.e. leave the light cone. But what is a "signal"? It is not sufficient to have a correlation between two events in order to have a signal. Reichenbach (1958) illustrates this with the following example : imagine a searchlight which projects light on a far-away surface (e.g. a cloud). If the searchlight rotates quickly enough, the velocity of the illuminated spot on the cloud may become larger than the speed of light, without any violation of the principles of relativity theory. The reason is that the spot of light cannot transfer information in its movement from, let us say, cloud A to cloud B, although there clearly is a correlation between the presence of a light spot in A and in B. Indeed, someone sitting at A cannot in any way modify or modulate the beam so that someone at B would be able to decode or interpret this modulation.

This analysis of signal transfer can be generalized to a definition of causality. According to Reichenbach (1958, p. 136), a relation between two situations, so that if the first situation is in class c, then the second situation is in class e, is causal, if a variation or modification of the first situation (so that c would be replaced by c*) is associated with a variation of the second situation (from e to e*). In other words, a *change* of the "cause" would lead to a change of the "effect". This is the same as requiring that the relation or process leading from cause to effect would form a "channel", which can be modulated so as to allow information transfer.

This definition can be easily assimilated to our distinction-based framework. Indeed, a *change* (from c to c*) implies the replacement of c by something (c*) which is *distinct* from c. This change or distinction ($c \neq c^*$) is then "transmitted" to the effect situation ($e \neq e^*$). However, Reichenbach adds to his definition the requirement that a variation in the effect (from e to e*) would not be associated with a variation in the cause (c). In other words, distinct effects may correspond to equal causes. Hence, this definition obeys only half of our causality principle (2), and corresponds to what we have called "reversible" processes,

which need not be predictable, however. This should not surprise us, since Reichenbach aims to define causality as potential *information transfer*, and reversibility is equivalent to the non-increase of entropy, i.e. the non-decrease of information.

The inclusion of the second clause (equivalent to unpredictability) in Reichenbach's definition is needed for establishing the antisymmetry of the causal relation. Indeed, the basis of the causal structure in relativity theory is temporal order, and this is derived from the order induced between events by the causal relation. Our definition (3), however, is in itself symmetric, and so we must make a more profound analysis if we wish to explain temporal order. Moreover, we would like to explain the existence of a limit speed (the velocity of light) for causal signals, which is assumed by Reichenbach and by relativity theory, but not explained (see section 5).

4. Partially causal relations.

We have shown by different examples that natural phenomena, as they are represented in (non-classical) physical theories, are in general incompletely causal. This means that distinction conservation, as required by (2) and represented by (4), is only valid for certain distinctions, but not for others. Furthermore, it is possible that the violation of distinction conservation only happens in one direction. Processes which are predictable but not reversible are characterized by the fact that distinctions cannot be created (i.e. distinct effects always result from distinct causes, cfr. (5)), but can be destroyed (i.e. distinct causes can have equal effects). Processes which are reversible but not predictable are characterized by the fact that distinctions cannot be destroyed (i.e. distinct causes always lead to distinct effects, cfr. (6)), but can be created (i.e. equal causes can have distinct effects).

We can extend this analysis to further types of "quasi-causal" relations. A causal relation is defined by two "connections" between classes (cfr. (3)):

$$c \rightarrow e, \text{ and } \tilde{c} \rightarrow \tilde{e}. \qquad (7)$$

An irreversible process can be defined by :

$$c \rightarrow e, \text{ and } \tilde{c} \rightarrow e \text{ (distinct causes, same effect e)} \qquad (8)$$

An unpredictable process then is characterized by :

$$c \rightarrow e, \text{ and } c \rightarrow \tilde{e} \text{ (same cause, distinct effects)} \qquad (9)$$

A different kind of relation, but which is often associated or compared with causality, is the *material implication* :

$$c \rightarrow e, \text{ and } \tilde{e} \rightarrow \tilde{c} \text{ (implication and its contraposition).} \qquad (10)$$

This relation clearly does not belong to the causality family, because cause and effect (c and e), are inverted in the second part of the definition, whereas causality is characterized by an unambiguous order between c and e. However, the confusion which sometimes arises between implication and causation can be understood by noticing that both are special cases of a weaker, i.e. more general property, which I will call *production* (the name comes from

the "production rules" which are used in artificial intelligence for modelling general inferences) :

$$c \rightarrow e \qquad (11)$$

Such a rule is to be interpreted as "if c, then e", without, however, presupposing anything about the case where c˜ or e˜ would be true. The fact that c and e must be distinguished from the situations in which they are not true in order to be meaningful, remains implicit. We can however explicitly add c˜ and e˜ in the formulation. If we consider the rule as expressing a weak form of causality (i.e. we explicitly assume that the contraposition $e˜ \rightarrow c˜$ is not valid), then we may assume that nothing is known in the case where c˜ would be true. In other words c˜ might lead as well to e as to e˜, and we get :

$$c \rightarrow e, \quad c˜ \rightarrow e, \quad \text{and} \quad c˜ \rightarrow e˜$$

This means that if we start with c, then the process is predictable. However, it is not reversible, since if we try to go back from e, we have a choice between two possibilities. If we start from c˜, the process is unpredictable, and irreversible if we find e. However, if we find e˜ to be the case, then we know that c˜ had to be the case, and we may reverse the process.

Such a production rule, which is partially irreversible and partially unpredictable, is clearly insufficient as a basis for constructing rational, scientific theories (cfr. Heylighen, 1989a). It offers rather limited control on the process it represents. However, it is the simplest type of connection between classes of phenomena, and as such can function as a primitive element by means of which more elaborate cognitive systems may be constructed.

After examining partially causal relations in the case of one distinction, we must study the case where several distinctions are involved. Since a set of distinctions determines a Boolean algebra, we may combine different distinguished classes by conjunction (to be denoted by "."). A causal relation may then be represented by a morphism of Boolean algebras (cfr. Halmos, 1974; Heylighen, 1987). A morphism is a mapping f between two algebras (which may be identical), such that the basic structures of the algebra are conserved. In the case of distinctions these basic algebraic structures correspond to the unary operation of negation, and the binary operation of conjunction.

The requirement (4) is then equivalent to negation conservation :

$$f : B \rightarrow B' : c \rightarrow f(c) = e \quad \text{is a morphism iff} \quad 1) \quad f(c˜) = (f(c))˜ = e˜$$

However, if we consider more than one distinction, we must add the conservation of the second operation which defines distinction (according to Spencer-Brown, 1969), conjunction (or equivalently disjunction), and we get the requirement :

$$2) \; f(c_1.c_2) = f(c_1).f(c_2) = e_1.e_2$$

This is to be interpreted as : if c_1 causes e_1, and c_2 causes e_2, then the conjunction of c_1 and c_2 causes the conjunction of e_1 and e_2. This requirement is again satisfied by the processes as they are represented in classical mechanics. Indeed each property or proposition describing a classical system can be represented as a subset of states ("atoms" of the Boolean lattice or algebra, cfr. Piron, 1976). Conjunction then corresponds to set-theoretic intersection, and this is conserved by the mappings which are used to represent (causal) processes in classical mechanics.

In the case of independent causes this requirement appears to be intuitively acceptable. For example, consider the propositions "rain causes the crops to grow" (i.e. "if it rains, the crops grow", "if it does not rain, the crops do not grow") and "wind causes the dust to be spread out". It seems natural to conclude that a conjunction of rain and wind causes the growth of the crops together with the diffusion of dust.

However, consider the proposition "a lighted match in the presence of an inflammable gas causes an explosion". Here we have a conjunction of causes (match and gas), whose effect (explosion) does not seem to correspond to a conjunction of the effects of the individual causes. To explain this we shall go back to our motivation for introducing the distinction concept. A distinction was said to be introduced in order to describe deviations from the "normal" situation. An explosion clearly is such a deviation. Therefore it is meaningful to introduce a distinction in order to discriminate between situations in which explosions occur and situations in which they do not. The presence or absence of a match together with a gas determines such a distinction. However, the presence of a match alone or of a gas alone does not produce a deviation from the "normal situation" (at least as far as explosions are concerned). Therefore we may conclude that the causal process leading to explosion must be represented by introducing a proper distinction (coincidence of lighted match and gas), and cannot be derived from causal processes involving the components (match and gas) of this distinction individually.

This example shows us again that an apparently causal connection (the conservation of distinction in one or the other direction) can in general not simply be extrapolated to other related distinctions, or to another direction (e.g. from predictability to reversibility). Causality appears to be a "local" concept, which can be used to describe certain aspects of certain processes, but which cannot be simply generalized in order to build an encompassing, deterministic and reversible theory, such as classical mechanics. In other words, general processes are only partially causal : they only conserve certain distinctions, perhaps only in one direction, and perhaps only during a limited time interval. We shall further examine why causality seems to work often but not always. But first we must study the relation between causality and time.

5. Causality and temporal order

One of the fundamental questions concerning causality is whether an effect must necessarily follow its cause. Or is time order merely an "accidental" feature of causal relationships, used to conventionally discriminate between the "cause" part and the "effect" part of an essentially symmetric relation? Distinction conservation as we have defined it, is a requirement which does not impose a specific order on the relata. But this does not entail that individual couples of causally related distinctions must be connected symmetrically. A priori, both possibilities, symmetric and antisymmetric (i.e. ordered, or oriented), can exist. However, if we are faced with a symmetric couple of relata:

$$\begin{array}{l} c \rightarrow e, \quad \tilde{c} \rightarrow \tilde{e} \quad \text{and} \\ e \rightarrow c, \quad \tilde{e} \rightarrow \tilde{c} \end{array} \tag{12}$$

then we will not interpret their relation as a causal relation, but as an equivalence relation: e if and only if c, $\tilde{e}$ if and only if $\tilde{c}$. This is to be interpreted that c cannot be actual without e

being actual, and c˜ cannot be actual without e˜ being actual. Remark that an equivalence thus defined (12) is both a causal relation (as defined by (7)) and an implication (as defined by (10)).

The "logical" relation of equivalence can be interpreted physically as a complete correlation between variables: the value of the one variable determines the value of the other one, and vice-versa. For example, in the EPR-paradox situation (Jammer, 1974) you know that if you measure spin-up for one photon you will find spin-down for the other photon, and vice-versa. The situation is completely symmetric. You will always find a correlation such as (12), whether you start with the left photon (c) or with the right photon (e), whether you find spin-up (c) or spin-down(c˜) (Heylighen, 1987).

The paradoxicality of the effect resides in the fact that one tends to assume that there is a causal influence: the measurement of one of the variables, leading e.g. to the actualization of c, is a quantum mechanical event which seems to cause another event, e.g. the actualization of e. However, the "influence travelling from c to e" is instantaneous, hence faster than the speed of the light, apparently contradicting the causality principle of relativity theory. However, if we do not interpret the effect as a causation, but as an equivalence or correlation, which differs only from classical correlations by the fact that the distinctions (c, c˜) and (e, e˜) are in a certain sense "created" during the quantum measurement process (cfr. section 3.), then there is no paradox in the instantaneous character of the effect (cfr. Heylighen, 1987). The theorem of Bell is not applicable because it implicitly assumes that all distinctions are conserved (i.e. that before the measurement they were merely "hidden" variables, already locally residing in the photons to be observed).

It seems natural to interpret the absence of (temporal) order in symmetric distinction conserving relations as the absence of duration between the actualization of the two distinctions. Hence equivalence is an "instantaneous" connection. Conversely we could then interpret the presence of order in an antisymmetric relation as an indication of a non-zero time interval between the two actualizations. However, time is a global property of a set of events and processes, not a local property restricted to one particular (causal) process. Hence we must examine the case were several distinctions and distinction relations are involved.

We want to construct time as a partial order relation on a network of events connected by causal processes. A partial order is characterized by antisymmetry and transitivity. By the argument above we may exclude all symmetric subrelations from the network, because these are not "causal" in the strict sense. We could then make the remaining network transitive by simple *transitive closure* (i.e. by adding all compositions of the relation with itself). This addition may be motivated by the fact that complete causality is transitive : if A causes B, and B causes C, then we may conclude that A causes C. Indeed, if all distinctions are conserved from A to B and from B to C then they are conserved from A to C. However, the addition of composed relations will create new symmetric subrelations: namely the compositions of "cyclic" subrelations (i.e. closed paths in the oriented network). If we want the network to become a partial order we will also have to exclude such cycles. This brings us to the problem of cyclic causality.

A *cyclic* causal relationship may be defined as a sequence of coupled cause-effect relationships which leads back to its starting point:

$$A \rightarrow B \rightarrow C \rightarrow \ldots \rightarrow A.$$

If we assume transitivity, we get : $A \rightarrow A$. There are several ways to interpret this formula in terms of distinctions. Let us begin with the case where A corresponds to one distinction

(a, a˜) (or where all distinctions represented by A behave in the same way). In that situation there are two possible interpretations. The simplest one is a tautology, "if a then a..." :

(a, a˜) → (a, a˜)

The transitive closure of the relation would make it into an equivalence relation between all the couples connected by the cyclic chain: (a, a˜), (b, b˜), (c, c˜) ...

Another possibility, however, of a chain of distinction conserving relations would be :

(a, a˜) → (b, b˜) → ... (a˜, a). (13)

Here, the order between a and a˜ is reversed. In other words, the causal chain sends a to a˜ (the negation of a), and a˜ to a. Let us assume first that the chain of processes has zero duration, i.e. all classes a, b, c, ..., a˜ are actualized simultaneously. In this case we would find a logical contradiction: a and not a would be simultaneously true. The chain could also be read as an equivalence : a if and only if not a, i.e. another contradiction. The only possible interpretation for such a proposition appears to be that there is no real distinction between a and a˜. They cannot be distinguished or separated: if you find the one, you simultaneously find the other one. It hence appears meaningless to allow formulas of the form (13) for describing causal (or logical) relations.

An example of the kind of problems arising from the literal interpretation of such inconsistent chains is the *paradox of the time machine* (cfr. Heylighen, 1987; Sjödin and Heylighen, 1985): if I would have a time machine, I could go back to the past and kill my father before he met my mother, so that I would not be born. But if I would not be born, I would be unable to kill my father and hence I would be born. In this case again, I would exist (a) if and only if I would not exist (a˜). The essence of time is just that time machines are impossible: it is only possible to move towards the future, not towards the past. The distinction-based model of formula (13) provides a very simple illustration of this fundamental property of time and causality.

We may conclude that cyclic chains of causal relations in which all distinctions are conserved in the same way have no temporal duration: either all elements of the chain are equivalent and hence simultaneous, or there is a contradiction due to the use of an ill-defined distinction, i.e. a distinction which entails its own converse.

But what about so-called *non-linear* (e.g. feedback, autocatalysis, ...) or *cyclic* causal processes (e.g. oscillations, periodical processes)? For example, if I become richer, I can make more investments, and hence I become even richer. If I become poorer, I must reduce my investments, so that I will gain less and hence become poorer. With symbols:

(richer, richer˜) → (investments, investments˜) → (richer, richer˜).

This is a form of positive feedback. Now a classical example of negative feedback, the thermostat: if the temperature goes down (d), the heating is activated by the thermostat (h), and hence the temperature goes up again (d˜). If the temperature becomes too high (d˜), the heating is interrupted (h˜), and the temperature goes down again:

(d, d˜) → (h, h˜) → (d˜, d).

Clearly, such processes cannot be interpreted as simultaneous correlations or as logical

contradictions. The reason is that there is an effective time interval between the first couple and the last couple of the chain of distinctions. But how can this be represented in the framework? Simply by marking the passage of time by an additional variable, i.e. a (set of) distinction(s). For example in the first case, the feature of getting richer is the same at the beginning and at the end of the causal chain, but the actual amount of money or property owned has changed, and can thus function as a measure of the fact that time has elapsed. In the second example, we could simply use a clock to detect changes between the beginning of the cycle (temperature going down) and the end of the cycle (temperature back at the same value and going down). The position of the hands of the clock will be distinct at these two instants, although the state of the thermostatic system may be the same.

In both cases a causal cycle can be characterized by the repetition of the value of one variable or distinction together with the change of the value of another variable, in contradistinction with the first two cases, where all values were either repeating or changing to their complement. In the second example, we could introduce the new distinction (t_n, t_n˜), meaning ("the time indicated by the clock is *n* or earlier", "the time indicated by the clock is later than *n*"). The cycle from temperature down (d) to temperature up back to temperature down could then be represented as:

$$(d, d\tilde{}) \rightarrow (d, d\tilde{})$$

However in parallel with this, another causal process, moving the hands of the clock, would be taking place:

$$(t_n, t\tilde{}_n) \rightarrow (t_{n+1}, t\tilde{}_{n+1}).$$

The conjunction of both determines a causal process which is partially cyclic, partially linear:

$$(t_n, t\tilde{}_n).(d, d\tilde{}) \rightarrow (t_{n+1}, t\tilde{}_{n+1}). (d, d\tilde{})$$

Let us summarize the properties of cyclic distinction conserving chains: 1. either all distinctions behave cyclically in the same way and then 1.1 the chain is part of an equivalence relation connecting simultaneous events or 1.2 it forms a logical contradiction, or 2. some distinctions cycle, but others change in such a way that the global state of the system is distinct from all preceding or subsequent states of the chain. The philosophy behind this analysis is that you can only have a time interval between two situations if something has changed between the situations, albeit only the state of the watch of the observer. If a distinguishable change has occurred, then at the global level (i.e. the level where all observed distinctions are taken into account) there was no real causal cycle, and it is possible to define an absolute ordering between all the events in the causal chain.

Let us conclude this section on time by a short reference to space-time as it is represented in relativity theory. Space-time is determined by its "*causal structure*" (Kronheimer & Penrose, 1967), consisting of a *chronological* relation (which is basically a partial order) and a *horismotical* relation. The union of both relation forms the "*causal precedence*" relation which is again a partial order. We have shown how a partial order, to be interpreted as causal precedence, can be constructed out of a network of causal connections. However, it is relatively simple to construct a horismotical relation too, by selecting acyclic causal chains which have no "parallel" chains (see Heylighen, 1987, for a formal proof). The chronological precedence can then be defined as the set-theoretic

difference between the causal precedence relation and the horismotical relation. In this way the essentials of relativity theory—in particular the existence of a limiting (i.e. horismotical) speed for causal processes—can be *derived* from the properties of distinction conserving relations, without any real additional assumptions.

6. The origin of causality

The last question we must address is : what is causality? Where does it come from? Is it an objective property of nature or merely a cognitive construct made by the observer?

Let us go back to the tautologous interpretation of the causality principle (1): *identic causes have identic effects*. This principle may be viewed as a description of what I would call *microscopic causality*. Let us illustrate this concept with the example of the radioactive nuclei. Macroscopically, i.e. at the level of the observer, the two nuclei are in all respects indistinguishable. Yet they decay at a distinct instant in time, thus violating the "macroscopic" causality principle (2). Microscopically, however, the two nuclei are not identic (otherwise there would be one nucleus, not two). Hence the fact that their decay times are not identic is in complete accord with the microscopic causality principle.

From the point of view of predictability, as we have already remarked, the microscopic causality principle is useless or trivial, since it ignores all repetition of processes or experiments. Predictability only exists on a macroscopic level, where microscopic differences between non-identic, individual phenomena are ignored, in order to determine the macroscopically meaningful distinctions between (infinite) *classes* of phenomena. Let us illustrate this basic difference between the two levels by looking at the discussion around determinism in physics.

It is a classic result that phenomena which are completely unpredictable at the microscopic level may be modelled macroscopically by deterministic theories. For example, statistical mechanics shows how the random collisions between molecules in a gas can be described by deterministic equations for macroscopic properties such as temperature, volume and pressure. It is also well-known that the indeterminism which quantum mechanics postulates for microscopic particles disappears when going to the "classical limit" of macroscopic objects. On the other hand, recent developments in self-organization models and non-linear thermodynamics have attracted attention to the opposite phenomenon: microscopically deterministic systems which behave in a completely unpredictable way when considered from a macroscopic view-point. Examples are the so-called "deterministic chaos", and certain types of cellular automata whose local dynamical rules are completely deterministic, but for which there is no global algorithm allowing to predict their overall evolution without computing all the individual, microscopic transitions from the given initial state (Wolfram, 1984).

We may conclude that in general there is no correspondence between the causal behaviour of macroscopic distinctions and of microscopic differences: in a given system, macroscopic distinctions may be conserved whereas microscopic ones are not, and vice versa.

This entails that the classical doctrine of *reductionism* (from macroscopic properties of systems to their microscopic constituents) is only useful in very limited cases: the causal relations between "emergent" macroscopic distinctions can in general not be derived from the eventual causal relations between their microscopic counterparts.

Another implication concerns the ontic interpretation of *determinism*. Until now we have used the word "determinism" in an epistemic sense: a theory was said to be

deterministic if all the distinctions made within that theory were causally conserved. Ontic determinism then would mean that you would infer from the determinism of your theory (e.g. classical mechanics) that the real world to which the theory referred were also deterministic (e.g. the world view of Laplace). In the present framework, the "theory" would correspond to the macroscopic distinctions made by the observer, the "world" to the microscopic differences between "non-identic" phenomena. From the above argument, it follows that distinction conservation and thus determinism cannot be transferred from the one level to the other one. On the other hand, you might postulate ontic determinism on the basis of the microscopic causality principle, but this principle is not only tautologous, it is also devoid of any operational significance. I would hence tend to conclude that the question of whether the world is (ontically) deterministic is a meaningless one, to which all possible answers are either trivial or unprovable in principle.

The discussion until now may have created the impression that distinctions, and hence causal relations, are purely epistemic entities, constructed by the observer. It is indeed the observer who discriminates between a finite number of classes, thus consciously or unconsciously ignoring an infinity of microscopic differences between phenomena belonging to the same macroscopic class. Yet the choice of which features to distinguish, although subjective, is not purely arbitrary. The purpose of distinguishing, indeed, is to be able to anticipate important changes in the environment, and hence distinctions will be selected on the base of their "predictive value", i.e. their "amount" of invariance or stability, in the sense of distinction conservation through causal relations. In other words, distinctions tend to be made if they "correspond" to dynamical regularities in the environment, characterized by (partial) invariances. Hence a distinction is neither purely subjective, nor purely objective.

But where do such external "regularities" come from? In other, words how do macroscopically conserved distinctions "emerge" from a microscopic world where distinction conservation is in the limit a meaningless principle? This emergence can be understood by introducing another tautological principle, which is not trivial, however, since some quite useful heuristic rules may be inferred from it. I am referring to the principle of *natural selection*, which in its most general sense states that stable systems tend to maintain (i.e. are naturally selected), whereas unstable ones tend to disappear (i.e. are naturally eliminated) (cfr. Heylighen, 1989b,c). A system here should in general not be understood as a concrete physical "thing" (e.g. a particle or a planet), but as a structure or organization with a recognizable identity. Such structures can in general be considered as sets of interrelated distinctions. "Stable" systems then are characterized by interrelated distinctions with some form of invariance (Heylighen, 1989b). Distinction conservation as defined in our analysis of causality is a very basic type of "invariance".

The emergence of causal relations, i.e. "natural laws", could then be understood as a process of natural self-organization, whereby microscopic systems would tend to group themselves in larger assemblies or classes by blind or random combination, and where *only those systems of classes would maintain where the dynamics of the system would conserve the distinctions between the classes*. A possible mechanism for analyzing this process might be found with the aid of the mathematical concept of "closure", as a criterion for determining invariant distinctions (see Heylighen, 1989b,c). Let me here just emphasize the self-organizing, evolutionary origin of causal laws.

One important consequence of such a view is that causal relations (i.e. conserved classes of distinctions) will be quite common, but not general. Natural selection is not an absolute, all-or-none mechanism, which only produces perfectly adapted, permanent systems. Hence we may expect that the conservation of distinctions will be only partial or

relative, that no system of distinctions would be absolutely stable. We should not be surprised to find that causal laws break down when we leave their domain of applicability (i.e. the domain of stability of the distinctions). Even fundamental "Laws of Nature" will break down in extreme circumstances such as those inside a black hole or during a "Big Bang" (cfr. Misner, Thorne & Wheeler, 1974). This explains our observation (in section 4) that distinctions are usually only conserved "locally".

If we believe that the universe has a beginning (e.g. the "Big Bang") then it is clear that causal laws must have arisen somewhere during or after that beginning: no causal law could govern the beginning itself, since this would entail that something distinct (the universe) would have been the effect of something without any form of distinctness (the primeval emptiness or Nothingness), in contradiction with our definition of causality as distinction conservation. All current theories of cosmic evolution assume that this evolution was characterized by a growing complexification, during which gradually more and more complex levels of systems have emerged: space and time, elementary particles, hydrogen atoms, more complex atoms, molecules, organic molecules, cells, organisms, human beings, societies, ... Each of these levels is characterized by its particular (causal) laws. Hence these laws must also have appeared gradually, one by one.

For example, the laws governing chemical reactions could not have appeared in a world in which no atoms or molecules existed, but only elementary particles. The concept of "valence" as an invariant distinction between molecules allowing to predict possible reactions is contentless without the sheer existence of molecules. Yet, once molecules could subsist as more or less stable assemblies of atoms, regularities or invariant patterns could emerge from their interactions. These invariant patterns of interactions could then form the substrate on which more complex patterns could self-organize, e.g. the whole of physiological processes allowing the survival of a cell. Much later, these same patterns of chemical reactions could be "discovered" by a human observer, who would introduce the concept of "valence" in order to facilitate the distinction between classes of molecules which reacted in clearly distinguishable ways.

A nice feature of the present analysis is that it allows to make a parallel between the "ontic" or "physical" emergence of a causal law and its "epistemic" or "cognitive" discovery by an observer. Both processes can indeed be analysed as forms of self-organization, either of a "physical" system or of a "cognitive" system attempting to model the behaviour of the physical system. The parallelism does not imply an isomorphism of the physical system and its model, however! The cognitive model will be necessarily infinitely less detailed, and will in general be equivocal, that is the same physical system will be represented in different ways, depending upon the point of view or the purpose of the observer (cfr. Heylighen, 1989b).

For example, an engineer designing a bridge will attempt to distinguish all factors causally related to the question whether the bridge will be able to carry the weight of the traffic passing over it. On the other hand he will usually not make a distinction between the factors determining the colour of the bridge. On the other hand, an artist planning to decorate the bridge will not be interested in discovering the properties of tension in the material, but will actively distinguish different colours in the components of the bridge, because he assumes that these are causally related to the general esthetic impression the bridge will make. Both observers, however, will only make distinctions which have some (causal) invariance; they will neglect all potential distinctions which are either so unstable that they cannot be observed or controlled (e.g. the quantum fluctuations of certain molecules in the material), or which have no causal relation whatsoever to any other relevant distinctions (e.g. the colour of the eyes of the politician who decided to build the bridge).

Thus, although their choice of what to distinguish is subjective, it is "objectively" constrained by the fact that a distinction must have a minimum of invariance in order to be meaningful.

7. Conclusion

I have presented a new framework for understanding and analysing causality, based on the introduction of distinctions as relata for the causal relation. A distinction is an implicitly defined primitive concept, relating a class to its complement, which can be thought of as having the formal properties of a Boolean variable. A distinction is selected by an observer from an infinite set of potential distinctions between non-identic phenomena. In order to be meaningful a distinction must have a minimal stability or invariance, i.e. during (certain) dynamical evolutions it must either remain constant or be mapped upon another distinction.

A process which maps distinctions bijectively (one-to-one) upon other distinctions is called causal. It is said to "conserve" distinctions, and is characterized by predictability and reversibility. If the mapping is not bijective but surjective (many-to-one) the process is irreversible, but predictable. If it is not bijective, but inversely surjective (one-to-many), the process is unpredictable, but reversible. In general processes are only partially causal: they only conserve certain distinctions, during a limited time interval, and often only in one direction.

The emergence of systems of (partially) conserved distinctions cannot be deduced from the properties of lower-level, "microscopic" distinctions, but may be understood as a process of self-organization, governed by variation and selective retention.

It was shown that the time ordering of cause and effect can be understood by analysing the properties of cyclic, i.e. not ordered, relations between distinctions. Such cyclic relations could be reduced either to order relations, or to equivalence relations (which in the strict sense are no causal relations), or they could be shown to be self-contradictory (causal paradoxes). It was argued that the resulting set of order relations can be interpreted as defining the "causal" structure of space-time, as it is defined in relativity theory.

The present theory has the advantage that it starts from a very simple and intuitively acceptable definition, yet allows to analyse the relations between very heterogeneous features of causality such as predictability, reversibility, cyclicity, space-time structure, observer-dependent and observer-independent aspects. As such, it provides a basis for tackling some of the classic paradoxes associated with causality in physics (see further Heylighen, 1987). Moreover, it proposes a first, tentative approach to the problem of the origin of causal laws.

It is clear, however, that a lot of questions remain to be answered. In particular, it is necessary to work out the different types and properties of, and interrelations between, partially causal relations (i.e. the general class). In order to understand in detail the emergence of globally conserved distinctions we must be able to model all the intermediate stages where distinctions have only limited, local invariances. In particular, it will probably be necessary to give up the assumption of the Boolean character of conserved distinctions, i.e. the conservation of conjunction and negation. This may allow us to build more detailed models of non-causal (and non-Boolean) theories such as quantum mechanics. Moreover such an analysis may give us more insight in the processes of discovery (or "induction") by which observers infer new conserved distinctions.

References

Einstein A., Podolsky B. & Rosen N. : 1935, Physical Review 47, p. 1804.

Halmos P.R. : 1974, Lectures on Boolean Algebras, Springer, New York.

Heylighen F. : 1987, Representation and Change. An Integrative Meta-representational Framework for the Foundations of Physical and Cognitive Science, Ph. D. Thesis, Vrije Universiteit Brussel, Brussels; to be published by Communication & Cognition, Gent.

Heylighen F. : 1988, Formulating the Problem of Problem-Formulation, in: Cybernetics and Systems '88, Trappl R. (ed.), Kluwer Academic Publishers, Dordrecht, p. 949-957

Heylighen F. : 1989a, Non-Rational Cognitive Processes as Changes of Distinctions, in: Proceedings of the International Congress C&C 20, F. Vandamme, M. Spoelders, G. Van de Vijver, M. Drolet & J. Van Dormael (ed.), Communication & Cognition, Gent.

Heylighen F. : 1989b, Building a Science of Complexity, Proceedings of the 1988 Annual Conference of the Cybernetics Society (London) .

Heylighen F. : 1989c, Coping with Complexity: concepts and principles for a support system, in: Preceedings of the Int. Conference "Support, Society and Culture: Mutual Uses of Cybernetics and Science", Glanville R. & de Zeeuw G. (eds.), OOC, Amsterdam.

Jammer M. : 1974, The Philosophy of Quantum Mechanics, Wiley, London.

Kronheimer E.H. & Penrose R. : 1967, On the Structure of Causal Spaces, Proceedings of the Cambridge Philosophical Society 63, p. 481.

Misner C.W., Thorne K.S. & Wheeler J.A. : 1974, Gravitation, Freeman, San Francisco.

Piron C. : 1976, Foundations of Quantum Physics, W.A. Benjamin, Menlo Park, California.

Prigogine I. : 1979, From Being to Becoming : Time and Complexity in the Natural Sciences, Freeman, San Francisco.

Reichenbach H. : 1958, The Philosophy of Space and Time, Dover, New York.

Sjödin T. & Heylighen F. : 1985, Tachyons Imply the Existence of a Privileged Frame, Lettere al Nuovo Cimento 44, p. 617 - 623.

Spencer Brown G. : 1969, Laws of Form, Allen & Unwin, London.

Wolfram S. : 1984, Universality and Complexity in Cellular Automata, Physica D 10, p. 1.