Causality in Complex Systems
An Inferentialist Proposal

Lorenzo Casini

University of Kent, Department of Philosophy

Submitted in accordance with the requirements for the degree of PhD

September 2012

Declaration

The candidate confirms that the work submitted is his own and that appropriate credit has been given where reference has been made to the work of others. This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

Abstract

I argue for an inferentialist account of the meaning of causal claims, which draws on the writings of Sellars and Brandom. The account is meant to be widely applicable. In this work, it is motivated and defended with reference to complex systems sciences, i.e., sciences that study the behaviour of systems with many components interacting at various levels of organisation (e.g. cells, brain, social groups).

Here are three, seemingly-uncontroversial platitudes about causality. (1) Causal relations are objective, mind-independent relations and, as such, analysable in objective, mind-independent terms. (2) There is a tight connection between our practice of predicting, explaining and controlling phenomena, and the use of causal notions. (3) The second platitude should be explained in terms of the first.

Contrary to this widely-held stance, I suggest that we reverse the order of analysis, by taking our activities of agents as the raw material in terms of which to account for the obtaining of causal relations. To this end, I propose and defend an inferentialist account of causality. Causality is a 'category' that the knowing subject employs to 'mediate' between himself and the world. In inferentialist terms, this mediation is the result of the concept of cause figuring in a network of inferences, used in our practice of gathering evidence and using it to explain, predict and intervene. Complexity only makes the mediation more difficult, thereby rendering the meaning of causality more evident.

Acknowledgements

I am deeply grateful to my supervisors Jon Williamson and David Corfield for their patience, advice and constant support.

I am indebted to Dominique Chu for introducing me to the topic of complexity, to Anjan Chakravartty for helpful discussions on dispositionalism and causality, to Ken Westphal for helping me find my way through Sellars' writings, and to Niels Skovgaard Olsen for extensive comments on chapters 7 and 8 and enriching discussions on inferentialism, normativity and objectivity.

I am grateful to Sylvia Nagl, Amos Folarin and May Yong from the Cancer Institute at UCL as well as Paolo Vineis and the EnviroGenoMarkers group at Imperial College London for letting me 'observe' scientists in action—and for allowing me to ask annoying questions about causality.

A special thank goes to Federica Russo and Phyllis McKay Illari for helpful conversations on causality as well as crucial tips on PhD survival.

I thank the School of European Culture and Languages and the Centre for Reasoning of the University of Kent for contributing to fund my PhD with, respectively, a teaching scholarship and a bursary for my editorial work for *The Reasoner*.

Research during my writing up year was conducted at Tilburg University and the University of Konstanz. I thank TiLPS and DAAD for generously funding these research visits. I am grateful to my mentors during these visits, Stephan Hartmann (TiLPS) and Wolfgang Spohn (Konstanz), for their feedback on, respectively, chapter 6 and chapter 7. I also wish to thank Wolfgang Spohn for the opportunity to work as a research fellow in the ANR-DFG co-funded project CAUSAPROBA, also in Konstanz. I am grateful to the people I came in contact with during these research stays for their helpful feedback on my research: Richard Dawid, Dominik Klein, Tobias Klein, Luca Moretti, Reinhard Muskens, Soroush Rafiee Rad, Pieter Ruys and Jan Sprenger (Tilburg); Michael Baumgartner, Jochen Briesen, Anna-Maria Eder, Eric Raidl, Jesus Zamora Bonilla and Alexandra Zinke (Konstanz).

I thank anyone who gave me his or her feedback at the conferences where my work was presented, among whom: Alex Broadbent, Nancy Cartwright,

Brendan Clarke, Francois Claveau, Carl Craver, Steven French, Alexander Gebharter, Donald Gillies, Stuart Glennan, Francis Heylighen, Chris Hitchcock, Andreas Hüttemann, Lars-Göran Johansson, Amir Ehsan Karbasizadeh, Kevin Kelly, Max Kistler, Samantha Kleinberg, Meinard Kuhlmann, Luis Mireles Flores, Alessio Moneta, Margaret Morrison, Steven Mumford, David Papineau, John Pemberton, Julian Reiss, Attilia Ruzzene, Gerard Schurz, Anders Strand, Mauricio Suárez, Micheal Tooley, Matthias Unterhuber, Erik Weber, Adam White, and Olaf Wolkenhauer.

Finally, I am very grateful to my examiners, Julien Murzi and Julian Reiss, for their valuable comments on the penultimate version of this thesis.

My sincere apologies go to all those that I forgot to mention but contributed somehow to this work. Needless to say, responsibility for any mistake or misinterpretation of ideas suggested by the people acknowledged here remains entirely mine.

Contents

Introduction

Causal claims belong to our ordinary language as well as scientific language. What is their meaning? Philosophers have been trying to answer this question since, at least, the times of Aristotle. But the topic is still a matter of controversy. This is all the more frustrating since we all think to grasp the concept of causality, and we tend to believe that by regimenting our intuitions on the concept it is possible to achieve a firm understanding of the nature of causality itself.

For instance, a widespread intuition, which one finds already in Hume's *Treatise*, is that a cause is something that makes a difference to the effect. Another, widespread intuition, which traces back to Aristotle himself, is that a cause is something that produces, or brings about, the effect. And we have other intuitions, too, e.g., that a cause is something that explains and is evidence for the effect, that the relation takes place in space and time, that it is asymmetric, local, transitive, etc.

Yet, in spite of repeated attempts to build solid analyses on these intuitions, the notion of causality continues to elude us. We keep debating on whether the causal relata are events, processes, properties, variables, facts, etc., on whether the causal connection is necessary or contingent, intrinsic or extrinsic, whether it involves the operation of powers or not, on whether it should be analysed in terms of general or singular facts, on whether it is an epistemic or an ontic notion, etc.

In this thesis, I purport to take a fresh look at the issue from a still unexplored point of view, viz. inferentialism. Inferentialism is a pragmatist position in semantics, leading to an unconventional sort of conceptual analysis. The motivation for this approach in the case of causality is the dissatisfaction with traditional approaches based on truth-conditionalist and representationalist semantics. The search for truth conditions or truth-makers has so far encountered problems on both conceptual and empirical grounds. On the one hand, for any account of causality on offer counterexamples can be found. On the other hand, no account seems to fully capture how the concept of causality is employed in scientific practice. Also, current approaches seem somehow

limited, or 'one-sided', insofar as they focus on *either* test conditions (which are then usually erected to truth conditions) *or* use conditions. This attitude reflects the one-sidedness of traditional theories of meaning, which focus on *either* conditions of application of expressions (e.g., verificationists, assertibilists, reliabilists) *or* consequences of application of expressions ("pragmatists of the classical sort") (cf. Brandom, 2000, pp. 63-66). Inferentialism, instead, gives equal weight to both components.

For the inferentialist, meaning is based on inferences, both intra-linguistic (language-to-language moves) and extra-linguistic (responses to non-linguistic circumstances and actions). The former are not to be interpreted only in descriptive terms, but also normative terms. Neither is to be interpreted in terms of reference or truth. The outcome of the analysis is not the identification of necessary and sufficient conditions, but the explication of the inferential connections that are constitutive of the meaning of sentences (e.g., causal claims) and subsentential locutions (e.g., 'causes').

It should be stressed that, since my main focus is the semantics of causal claims, I won't be directly concerned with metaphysical issues. For the inferentialist, causal claims are analysed in terms of endorsement conditions, not truth conditions. This does not entail, at least not straightforwardly so, any metaphysical stand on the nature of the relation that is supposed to make the claim true. However, some qualifications are needed.

First, inferentialism does allow one to re-interpret traditional issues such as whether causal talk is referential and in virtue of what. Reference is, for the inferentialist, not what grounds meaning but a consequence of the social and normative attitudes that institute meaning. It is only in this circumscribed sense that, in this thesis, semantics will serve the purpose of drawing conclusions on the nature of the relation talked about in the claim.

Secondly, for the inferentialist metaphysical investigations require a study of the jobs that concepts do for us, e.g. of the explanatory function of the verb 'causes'. It is usually thought that an understanding of the metaphysics of causation is sufficient to deliver an account of causal explanation. This is problematic, since the asymmetry of explanation is not reducible to the asymmetry of causation: why do causes explain effects and not the other way round? A better approach is to take the epistemology of causation to provide an account of causal explanation: in fact, the observation that the cause makes a difference to the effect *is* conducive to causal explanation. Still, one may ask: why are causal epistemology and causal explanation both

causal? My suggestion is to use inferentialism to understand both 'causes' and 'causally explains'. There will be inferences constitutive of the meaning of causal claims and inferences constitutive of the meaning of causal explanations. Here I focus on the former.

The account I propose is meant to be widely applicable. In the thesis, it is illustrated and defended with reference to complex systems. Why complex systems? Because complexity calls into doubt our intuitions about causality and makes causal talk problematic. So, complex systems make a better test case to uncover the meaning of causal claims. Arguably, if their meaning can be explicated in complex systems, it can also be explicated in areas of discourse where causal talk is less or equally problematic. In particular, I will mainly refer to two case studies, one from systems biology (apoptosis), the other from computational economics (asset pricing). This choice is meant to help develop an account that is encompassing enough as to bridge the natural-social divide, so that the account can be more easily applied outside complex systems. Why complex case studies rather than toy examples? There are several reasons for this choice. First, rather than toy examples, real science examples serve better the *desideratum* that a *scientific* account of causality be faithful to the notion of causality at work in scientific *practice*. Secondly, contrary to the intuition that different analyses are best suited to different areas of inquiry, no traditional analysis seems particularly well suited to complex systems sciences. Finally, complex systems are also convenient because a handful of examples are sufficient to drive my point home.

The thesis is structured as follows. In chapter 1, I give an informal characterisation of complexity. In chapter 2, I motivate my project by arguing for the compatibility of complex and causal phenomena and for the need of a suitable interpretation of causality in complex systems. In chapter 3, I introduce the reader to the case studies and argue for the causal interpretability of the models provided by complex systems sciences. With reference to this picture, I discuss and criticise monistic accounts in chapters 4 and 5, and pluralist accounts in chapter 6. In chapter 7, I develop my inferentialist account on the meaning of 'causes', which I illustrate with reference to complex systems in chapter 8.

Chapter 1

The Advent of Complexity

Causality has always been a hotly debated topic in philosophy. Complexity, on the other hand, is an equally hot but relatively new topic of philosophical discussion. Before embarking on the arduous task of discussing causality in complex systems, it is appropriate to characterise, even if only informally, the notion of complexity. Since 'complexity' is hard to define (§1.1), I will proceed by first defining 'simplicity' (§1.2) and then adding complications until *prima facie* genuinely complex phenomena emerge (§1.3).

1.1 What is complexity?

The last thirty years have witnessed an increasing interest within many scientific fields towards the phenomenon of complexity. This increase has occurred for several reasons. On the one hand, there has been a feeling of widespread dissatisfaction with the 'traditional' scientific approach—i.e., an approach that postulates idealised physical systems and perfectly rational agents to reduce hard tasks to simpler ones, analysable in terms of stable equilibria, linear relations, periodic motions, controllable variables, etc. This traditional approach, in fact, leaves untouched pressing questions such as: How does living matter arise out of *non*-living matter? Why do financial crises occur *so* often? On the other hand, this increasing interest towards complexity depends on the success with which new mathematical models and tools, originally developed to deal with a handful of striking, peculiar cases (weather forecasts, fluid turbulence, population dynamics, etc.), have been applied to a large variety of phenomena. Complexity has been advertised as a novel, revolutionary, 'post-Newtonian' paradigm in opposition to classical, 'Newtonian' science. This has motivated the hope that a new science with its own subject matter was about to spring forth.

Despite this growing interest, however, the existing variety of approaches

that fall under the label of "complexity science" still hasn't received a unified and generally accepted treatment. This is because there is a substantial disagreement and ambiguity on what 'complex' means, and how complexity can be defined, measured, etc. Existing boundaries between complex and simple phenomena are vague, and available criteria to distinguish between complex and simple are not unanimously accepted. It is important to stress that defining complexity is not the task of this thesis. For my purposes, it will be sufficient to give an intuitive, non-formal characterisation of systems that are commonly regarded as complex, and then go on and show what notion of causality is best applicable to them.

There exist formal *measures* of complexity, viz. *algorithmic* and *statistical* complexity, which count the number of symbols of the shortest program that produces the data (algorithmic) or statistically significant features of the data (statistical). However, the two measures can deliver conflicting verdicts with regard to the same situations. For instance, a maximum entropy situation has maximum algorithmic complexity but minimum statistical complexity—it is complex to reproduce exactly position and momentum of the particles, but simple to describe statistically significant features, such as their mean and variance. It's hard to decide between these or other measures in the absence of a notion of what it is exactly that they are supposed to measure.

Many *definitions* of complexity, whether formal or informal, have been provided so far, none of which has, however, been unanimously endorsed.

There have been attempts to provide necessary and sufficient conditions for complexity. For instance, Rosen has proposed that a system is complex if and only if some of its models are non-Turing computable (see Rosen, 1998, p. 292). It is debatable, however, whether some formal feature of the models should be taken as bearing on the ontology of the systems that such models are meant to represent.

Most definitions of complexity have an informal character. Some of them tend to focus on objective features of complex systems such as self-organisation ('order through fluctuations' (Prigogine and Stengers, 1984), 'self-organised criticality' (Bak, 1997), 'intermittent criticality' (Sornette, 2002)), adaptation (Holland, 1995), 'autopoiesis' (Varela et al., 1974), or some mixture of these (see, e.g., Mitchell, 2003, Part I). Other characterisations, instead, make also reference to subjective aspects, such as observer-relativity (Gersherson, 2002) or 'contextuality' (Chu et al., 2003). As we shall see (§1.3.4), the factors responsible for complexity can be highly contextual, i.e., their salience changes

from situation from situation.

Given the current unavailability of unambiguous and unanimously accepted definitions of complexity, I will proceed towards a characterisation of complex systems suitable for my investigation in a piecemeal fashion, by first characterising 'simple' systems and then adding 'complications' until *prima facie* complex features appear. This will help me illustrate in what sense, and to what extent, complexity has modified our notion of causality.

1.2 Simplicity

The notion of simplicity is usually associated with classical, 'Newtonian' science (see Prigogine and Stengers, 1984, chap. 2). Newtonian science was guided by the belief that every event, or state of a system, is determined by some previous conditions together with dynamical laws that specify the evolution of the system's behaviour in time. This opened the way to the idea that the universe is made of machines, and is itself a big machine, such that its behaviour can be in principle predicted, controlled, and exploited by knowing the laws of its working. It is no coincidence that after the rise of classical mechanics, which purports to describe the relative motions of massive bodies by reference to just four general principles, viz. Newton's laws, the myth also arises of an omniscient being, or 'demon', that in virtue of its knowledge of position and momentum (i.e., the product of mass and velocity of a body) of all bodies in the universe at a given time is able to infer the state of the universe at any other time, whether in the past or in the future (see Laplace, 1902, p. 4).

In classical mechanics, the trajectories of the particles of any closed system (i.e., a system that does not exchange energy or matter with the environment) can be, in principle, fully specified. Such trajectories are (i) *lawful*, (ii) *deterministic*, and (iii) *reversible* (see Prigogine and Stengers, 1984, p. 60). Once positions and velocities of the bodies at some time are known, together with the equations of motion that relate the dynamic forces to which the bodies are subjected to their acceleration, other states of the system are lawfully deducible, fully determined, and reversible by an external intervention consisting in a velocity inversion of all the bodies, which makes the system go 'backward in time' through its previous states and restore its initial conditions. This inversion is practically impossible to perform precisely in reality, due to the non-negligible external forces which make the system open—not

to mention the difficulty involved in dealing with huge numbers of particles. But if the universe itself is a closed system (the closed system par excellence), then nothing prevents an extremely powerful being such as Laplace's demon from rewinding the tape.

What is causality within this classical framework? The notion of causality overlaps greatly with that of *determination*, or *necessitation*. Metaphysically speaking, causal relations are relations between states of the system, such that each state necessitates the state that follows and is necessitated by the state that precedes along a continuous temporal chain. Under the Galilean assumption that the book of Nature is written by God in *simple* mathematical terms, such relations can be, in principle, known and exploited for explanation (retrodiction provides a kind of explanation, viz. *ætiological* explanation), prediction and intervention (e.g., velocity inversion) (see Israel, 2005).

This reasoning seems to rely on the assumption that legitimate causal talk depends on there being a metaphysical commensurability, or resemblance, between cause-and-effect events based on the permanence of the substance present in each (see Descartes, 1996, p. 28).[1] To make sense of the commensurability of the states in the determination relation, it must be possible in principle to adequately describe cause and effect in terms of their intrinsic properties and to specify the temporal evolution of the system in terms of universal, deterministic laws. Originally, what had to be commensurable was matter. Now the view is that it is the complex energy-matter that must be conserved (see §5.4.4). However, the same reasoning applies to both cases.

It must be stressed that in this classical framework the metaphysical and the epistemological aspect are reconciled in the figure of Laplace's demon. The demon, in fact, represents an ideal of absolute objectivity, of perfect knowledge of mind-independent, metaphysical relations. And although such a perfect knowledge is infinitely distant from the partial knowledge of limited beings like us, according to the classical paradigm this is for practical not theoretical reasons: our imperfect knowledge—e.g., of causal relations—can always be *perfected*; absolute objectivity—e.g., of causal claims—can be progressively *approximated*. Now, after three centuries of Newtonian science, this ideal of lawfulness, determinism and reversibility appears at least problematic. A unbridgeable chasm seems to separate us from Laplace's demon.

I will now illustrate how the introduction of 'complications' modifies the behaviour of simple systems, thereby affecting this classical picture. Notice

[1] For a criticism of this view, see Goldstein (1996, §3.2) and references therein.

that such complications may all be given an ontological interpretation. However, in the present context, due to my methodological choice of investigating causality by looking at models of complex systems, more important are their epistemic consequences, i.e., the way complexity affects the models' descriptive adequacy, their capacity to function as tools for surrogative reasoning about their targets, etc.

1.3 Complications

1.3.1 Nonlinearity

Since Newton's times, the problem of determining analytically (i.e., by integration of the differential equations specifying motion) the trajectory of a system composed of more than two bodies mutually attracting each other has proved extremely hard to solve. The original problem, the 'three-body problem', consisted in calculating the influence of the Sun on the Moon's motion around the Earth. This was later generalised to n-body systems where three or more bodies interact with one another at whatever size level, also microscopic, and where the nature of the interaction is not limited to gravitational forces. The apparent impossibility to solve the equations analytically, that is, without relying on numerical approaches, calls into question the first of the classical assumptions, that is, the idea that dynamics are always lawfully deducible. The problem, it is often said, resides in the 'nonlinear' character of the system. As we shall see, nonlinearity is a necessary—but not sufficient—condition for complexity.

Dynamical systems theory represents all possible dynamics in the *phase space*. In it, each possible state of the system corresponds to a point. For mechanical systems such as n-body systems, the phase space usually consists of all possible values of position and momentum. More generally, the phase space can be used to represent the evolution of *any* variable as a function of time, e.g., temperature, pressure, concentrations of chemicals in a reactor, gene frequencies in population genetics, electrical activities of neurons, populations of different species in an ecosystem, etc.

The dynamics of the system, whether linear or nonlinear, is usually described by using differential equations, which represent the evolution of systems in continuous time.[2] Among these, one can distinguish between ordinary

[2] Alternatively, difference equations, or 'iterated maps', are employed, which describe the evolution of the system in discrete time steps (see §1.3.3).

differential equations (ODEs), which have only time as independent variable, and partial differential equations (PDEs), which have also other independent variables, e.g., one or more spatial dimensions.

Linear differential equations can be expressed as a linear combination of the state variables. They are particularly important because, for them, *superposition* holds. According to the principle of superposition, if an equation is linear, the sum of any two solutions is always a solution:

$$F(x_1 + x_2 + ... + x_n) = F(x_1) + F(x_2) + ... + F(x_n) \tag{1.1}$$

The superposition principle, however, does not generally hold when the equations are nonlinear. In this sense, the complexity of the three-body problem consists in the impossibility to reduce motion of the entire system to the motions of the three bodies in superposition with one another, i.e., to reduce the differential equation describing the motion of the three bodies jointly considered to a—solvable—system of linear equations. This is regarded as one of the paradigmatic examples of the widespread non-reducibility of whole systems' behaviours to the behaviours of their parts studied in isolation (see Casti, 1994, pp. 40-41).[3] However, this does not seem sufficient to characterise complexity for at least two reasons.

First, in certain (though, admittedly, few) cases it is possible to find analytic solutions to nonlinear equations.[4] What are, then, the 'nice' nonlinearities and what the 'nasty' ones? There are no sharp boundaries. Besides, why should we take the (non-)existence of exact solutions as bearing on how much nature is complex in itself? The problem seems to reside more in our ability to solve the equations than in some blueprint for complexity in nature. As we shall see, qualitative methods show that the *general* behaviour of nonlinear systems is sometimes very simple (e.g. a pendulum) and *deterministic* (read: 'fixed' by the dynamical laws), although *specific* states are often *non-determinable* (read: 'non-analytically deducible' from those laws). So, often one must rely on numerical analysis or simply wait that the system unfolds

[3] 'Stability'—i.e. insensitivity of the planetary orbits to small perturbations, which prevents planets from colliding, on the one hand, and flying off in space, on the other—is a property of the whole which 'emerges' from nonlinearities and is irreducible to the parts' behaviour. For more on emergence and its connections with causality, see §2.2.

[4] For instance, the superposition principle does not generally apply to nonlinear waves. However, although there is no general analytical method to find the solutions of the nonlinear equations describing such waves, some of these equations are still solvable, e.g., the Korteweg-de Vries equation, a third-order (partial) differential equation describing small amplitude, shallow water waves.

itself to know what has/had to happen.

Secondly, and more generally, most real systems are nonlinear.[5] In fact, linear systems are typically at or near equilibrium, whereas nonlinear systems are far-from-equilibrium (see §1.3.3), and far-from-equilibrium systems are much more common than equilibrium ones. So, if nonlinearity were taken as the defining characteristic of the complex, almost anything in nature would count as complex, with the result that the term "complex" itself would not stand for anything special and worth calling "complex".

Other ingredients must be added to nonlinearity to get complex behaviours.

1.3.2 Sensitivity to initial conditions

The interstellar three-body problem involves a 'conservative', or nearly conservative, system. However, most of the systems displaying complex behaviour are non-conservative, or 'dissipative'.

A conservative system is such that its total energy is not dissipated. An example is an ideal pendulum not subject to friction. In order to find analytic solutions of the equation describing its motion, textbooks usually reduce the problem to the study of a 'simple harmonic oscillator', that is, a pendulum subject to a force *exactly* proportional to the angle of displacement from its resting equilibrium position. This approximation gives accurate results when the angle is sufficiently small. Although for large angles the maths gets complicated, topology, viz. a branch of mathematics, can still describe the qualitative behaviour of the pendulum for *any* angle of displacement in relatively simple terms. In figure 1.1 (top) are represented both the oscillatory behaviour of the pendulum (along the circles) and the rotatory behaviour (along the sinusoids). Both behaviours are periodic.

Now, let us add friction so that the pendulum becomes non-conservative (bottom of figure 1.1). A rotating pendulum will slow down more and more, 'passing' continuously from one energy level (a line in the phase portrait of the conservative pendulum) to the other, until at a given point its motion will become oscillatory and, eventually, stop. Have we reached the most that complexity has to offer? Not quite.

In the plane, the only two *attractors* for the behaviour of non-conservative systems—that is, the *stable limits* the system can reach as time advances—are *fixed points* (the motion of the pendulum subject to friction terminates

[5] See examples in (Strogatz, 1994), (Klipp et al., 2009) and (Gilbert and Troitzsch, 2005).

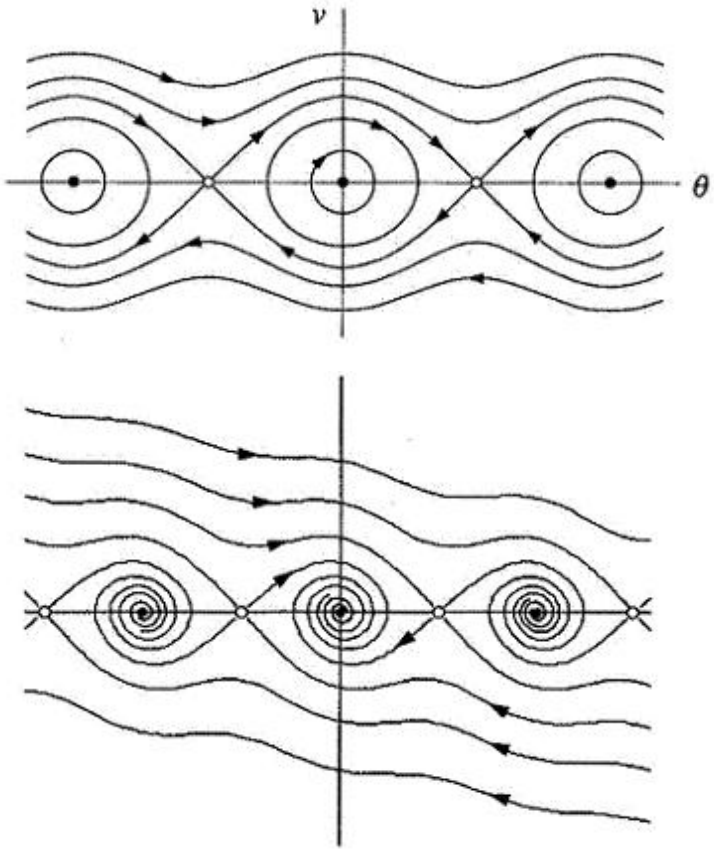

Figure 1.1: Phase portraits for the motion of a conservative pendulum (top) and non-conservative pendulum (bottom). Reproduced with permission from (Strogatz, 1994, pp. 170, 173). For systems whose motion can be described by means of one space coordinate only, such as the angle of rotation θ of a pendulum, all possible trajectories can be represented in a two-dimensional phase space. In the case of the pendulum, the two state variables will be θ and $v = \dot{\theta}$.

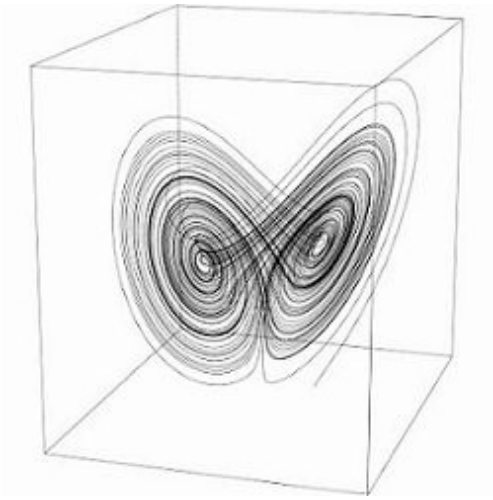

Figure 1.2: The Lorenz attractor is—for a suitable choice of the parameters σ, ρ and β—the limiting behaviour of the following, nonlinear system of differential equations:

$$\begin{aligned} dx/dt &= -\sigma x + \sigma y \\ dy/dt &= -xz + \rho x - y \\ dz/dt &= xy - \beta z \end{aligned}$$

here; see bottom of figure 1.1) and *limit cycles* (which attract the trajectories nearby the cycle, both inside and outside; this is the case of, e.g., pendula with an escapement mechanism, which neutralises the effect of dissipative forces). There is a sense in which, although the quantitative behaviour of a pendulum may be hard to calculate exactly, it is also very simple: it is either a fixed point or a limit cycle.

In more-than-two dimensional spaces, instead, there exists another kind of attractor for continuous dynamical systems, viz. the 'strange' attractor. Strange attractors constitute the limiting behaviour of *chaotic* systems.[6] I briefly illustrate the properties of chaotic systems with reference to the toy model of the weather that Lorenz offered to suggest the impossibility of long-term weather forecasts (see Lorenz, 1993, p. 188). For some suitable choice of the parameters, the system converges to a 'strange' attractor (see figure 1.2).

In general, a strange attractor is *structurally stable* (small perturbations

[6]Notice that very few systems are in a chaotic state, although they can be driven to chaos by fine-tuning some control parameter. For instance, the Lorenz attractor arises only for some values of the parameters not others (see below). Instead, most complex systems 'spontaneously' move from one stable configuration to another, organising themselves 'at the edge of chaos' (see §1.3.3).

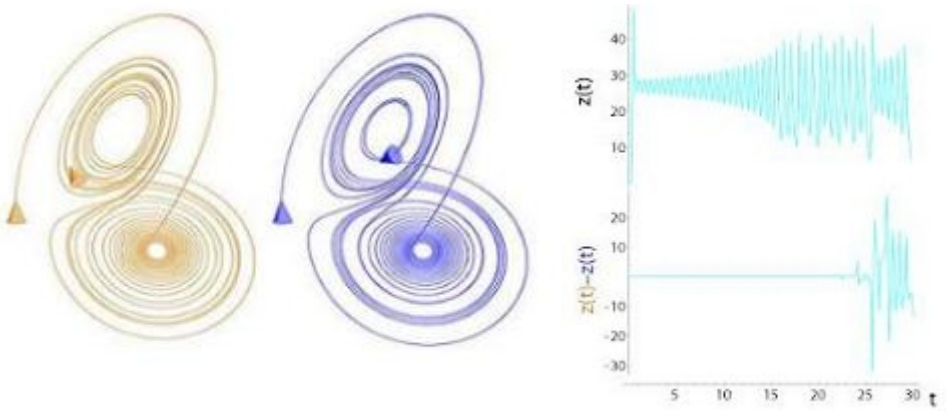

Figure 1.3: Sensitivity to initial conditions in 3-d and time-series. Left: two Lorenz orbits for $0 \leq t \leq 30$. The red orbit is one trajectory of the Lorenz system with $(\sigma, \rho, \beta) = (10, 28, 8/3)$ started from the initial point (0, 0, 1). The blue orbit is for the same parameters but initial condition $(0, 0, 1 + \varepsilon)$, $\varepsilon = 10^{-5}$. Right: the time-series for the z-coordinate of the first orbit and the difference between this and the z-coordinate of the second orbit.

do not change its topology), *aperiodic* (no point comes back exactly to where it was before), and *fractal* (its orbits are infinitely long but contained in a finite area, the dimension of which is not an integer, e.g., 2.5 or 3.12).

Strange attractors are such that any two points as close as possible to one another—but not coinciding—belong to trajectories that diverge at an exponential rate, that is, more than linearly. This means that chaotic systems are extremely sensitive to initial conditions, and display the so-called 'butterfly effect': small differences in initial conditions, huge differences in distant future states (see figure 1.3).[7] Extreme sensitivity to initial conditions entails that complexity can be the result of very simple processes, and that systems governed by a few, simple, *deterministic* equations can generate *unpredictable*—hence, complex—behaviour.

Notice, however, that despite this extreme sensitivity there are still margins for short-term (weather forecasts) and qualitative (climate change) pre-

[7] A special case of chaotic systems are systems where the basins of attractions, rather than the attractors themselves, are *fractal* (see Prigogine and Stengers, 1984, pp. 263-264). Here, the impossibility to specify the effect of an infinitely small perturbation on a point nearby the boundary between the two basins depends on the impossibility to specify, by increasing the resolution to any less-than-infinite degree, whether the point lies in one or the other basin, that is, whether it belongs to a trajectory that terminates in one or the other attractor.

dictions. For any two states sufficiently similar to one another, the evolution of the system will be sufficiently similar for a while, before the two trajectories diverge. The idea is that, if today's weather is sufficiently similar to some other day's weather in one's records, then it is likely that today's weather will evolve as the past weather did. This intuition is exploited in several ways in forecasting practice. For instance, one may look for close similarity not between two single states of the system—which one cannot compare in the absence of a good mathematical model of the system—but between two stretches of a time-series, called 'motifs' (see Stewart, 1997, p. 131). In a time-series, the state of the system is monitored by regularly plotting the value of one of its variables along one axis against the time of the measurement, as time progresses, along another axis (see figure 1.3, right). If the points in the time-series belong to the same, chaotic, attractor, then one can concentrate on the qualitative 'texture' of the time-series. Observation of two similar, recurring patterns of points at different times is an indicator that the system is chaotic and the trajectories to which the two patterns belong will be similar in the short-run. If, instead, one does have a model, one can proceed by comparing not just single points in the phase space, but groups of points and their relative distances, by using a method called 'tessellation' (see Stewart, 1997, pp. 292-295). The idea is that the forward state A' of A will bear to states B', C', D', etc. which lie in the forward trajectory of A's neighbour states B, C, D, etc., the same relation that A bears to B, C, D, etc.

Although complexity and chaos are tightly related notions, there is more to complexity than chaos. Let us complicate the picture a bit more, then.

1.3.3 Bifurcations and symmetry-breaking

Many systems in nature organise themselves into stable configurations that are open to the environment enough for them to sustain their autonomy but not to the point of losing their organisation. In general, the mechanisms that enable systems to reach these organised states are taken to be the root not just of complexity but of life itself (with which complexity is often associated), as they bring about highly differentiated, self-organised and, in some cases, goal-oriented forms of behaviour. This idea, in turn, is meant to constitute a crucial shift from traditional thinking, namely from the dogma of reversibility, which cannot account for how the living arises out of the non-living.

Whereas classical mechanics envisages the states of the system and time it-

self as reversible, thermodynamics interprets them as irreversible. Also closed systems, that is, systems that do not exchange energy with the outside world, but for which temperature is allowed to vary, despite conserving their total energy, do lose their ability to use this energy to do work because of friction and heat loss, which results in entropy increase. Entropy can be informally defined as a measure of disorder of a system, that is, the degree of randomness in the arrangement of its particles. The universe, however, in patent opposition to the idea that progress is (just) towards randomness and disorder, organises itself more and more in highly ordered structures. Paradigmatic examples of complexity, such as the cells' metabolism, are forms of organisation that arise when open systems are far-from-equilibrium, or 'at the edge of chaos'. In virtue of what is this process of organisation possible?

The idea is that dissipative systems, that is, open systems that promote overall entropy increase, can organise themselves at the expense of the environment, by using the energy and matter that flow into them to do work and by getting rid of the excess of entropy by exporting it out of the system in the form of waste products. In this way, they reduce their internal entropy, i.e. increase their organisation, by increasing the entropy of the environment. Given that many systems have more than one attractor, this process also contemplates a continuous adjustment to changes in boundary conditions. The resulting self-organisation mechanism need not involve centralised control but can be, in a sense, 'spontaneous'. It usually starts with a positive feedback, such that an initial fluctuation is amplified. Then, after a stable configuration has emerged, the forces that at first reinforced the positive feedback act in a negative feedback so that deviations from stability are suppressed. This phenomenon was observed in many natural systems, from convection cells to chemical and biochemical systems (e.g. chemical clocks, cells, the brain) to ecological systems, economic systems, etc. I will first present an example where 'fine tuning' is required for self-organisation, and then generalise to cases of spontaneous self-organisation.

Let us consider convection cells, first observed by Henri Bénard. Convection is a very general phenomenon that takes place in a large variety of out-of-equibrium systems, in liquid or gaseous state, at small and large scales. It can be reproduced by first heating the liquid from below and letting it cool on the top, and then progressively increasing the temperature at the bottom. As a result, the lower part becomes less dense and tends to rise towards the top and, at the same time, the cool liquid at the surface tends to sink towards

the bottom. Convection arises as these two movements become coordinated so that hexagonal cells, or 'rolls', form, with an upward flow on one side of each cell and a downward flow on the other side.

Notice that the formation of convection cells is, in a way, fully deterministic, as it can be predicted by knowing dynamical laws and temperature difference. However, whether liquid molecules in a given region will decide to move up or down cannot be accounted for in deterministic terms. *Symmetry-breaking* involves a 'choice' between two possibilities, corresponding to two different stable configurations. This choice is unpredictable: if one repeats the experiment many times, whether a particular region will be in a clockwise cell or in a counter-clockwise cell is a statistical affair. This mechanism can be represented by means of a 'bifurcation diagram'. A bifurcation diagram (see, e.g., bottom of figure 1.4) shows how passage from one stable configuration (e.g., rest) to another (e.g., a periodic oscillation) depends on two factors: (i) a change in some control parameter (e.g., temperature) measuring the distance from thermodynamic equilibrium and (ii) a choice between two new, far-from-equilibrium configurations (e.g. clockwise or counter-clockwise configuration). The temperature at which convection appears corresponds to a *bifurcation point*. Further increase in the control parameter determines more and more complex spatio-temporal configurations. Other bifurcations occur, namely, a succession of period-doublings: first a period-two oscillation (one maximum and one minimum), followed by a period-four (two maxima and two minima), a period-eight, and so on until a totally chaotic, aperiodic, regime is reached as the parameter approaches a critical value. This feature does not just apply to convection cells but is shared by all self-organising systems whose dynamics depend on some control parameter. Also, the ratio of convergence between parameter values corresponding to successive period-doublings is constant among large classes of phenomena. This important property, originated in the study of phase transition phenomena in statistical mechanics and later discovered by Feigenbaum in simple 'iterated maps', is called 'universality'.

An iterated map is a map representing the behaviour of an iterated function, that is, a function that successively maps values of x onto other values of x through composition with itself, so that $f^0(x) = x, f^1(x) = f(f^0(x)), f^2(x) = f(f^1(x)), ..., f^n(x) = f(f^{n-1}(x))$, etc. Feigenbaum started by examining the (discrete) logistic equation (see top of figure 1.4):

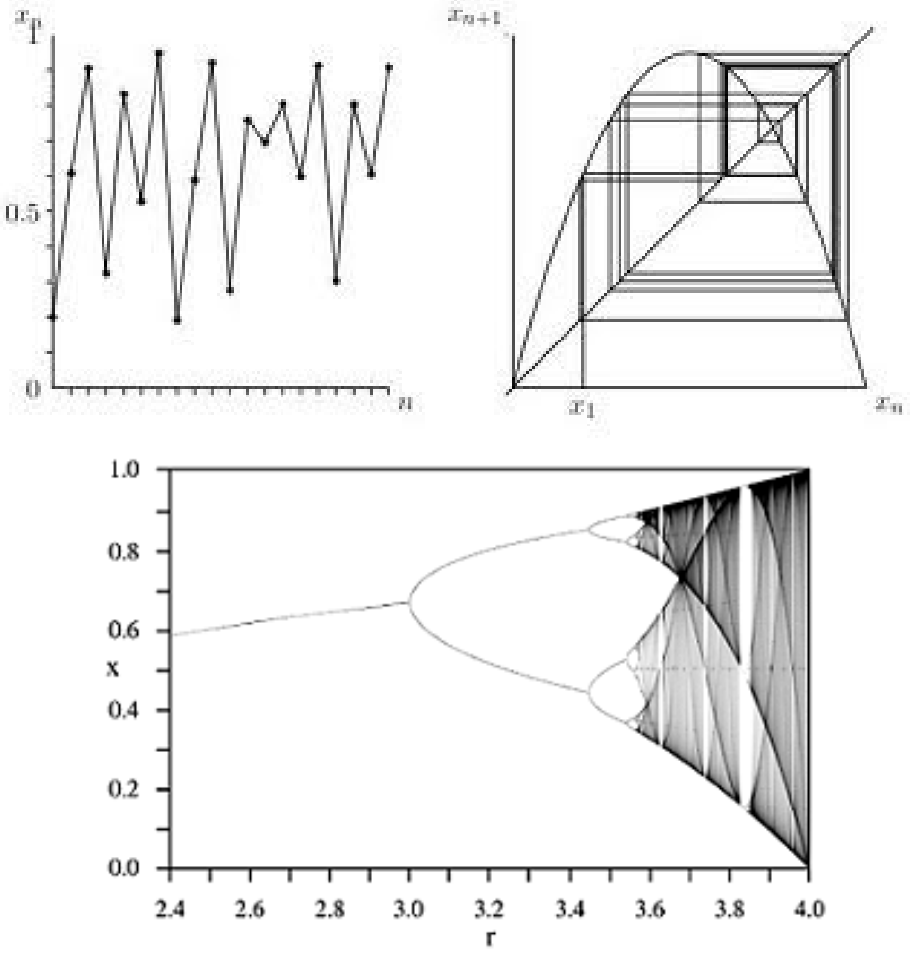

Figure 1.4: Logistic map (top right) and bifurcation diagram (bottom) for the logistic equation. Iterated maps are used to describe the behaviour of the same function depending on the value of some control parameter. As the parameter r changes, the system first settles into a fixed point then goes through a period-doubling cascade until it reaches chaos.

$$x_{n+1} = rx_n(1 - x_n) \tag{1.2}$$

which is usually employed to represent the growth of some quantity (e.g., this year's population with respect to last year) depending on the availability of some resource (e.g., the carrying capacity of the environment) and he found a certain convergence ratio. Then he moved on to many other difference equations and concluded that *all* functions (provided they all had a one-humped mapping and the hump resembled a parabola) shared the same convergence ratio, which is a *universal* property. Universal properties can be shared by complex systems governed by different dynamical laws, irrespective of the microscopic details of their constituents, and may be, to varying extents, used for prediction, control and explanation.

Phenomena of self-organisation, produced and maintained by interlocking positive and negative feedback loops, and involving bifurcations, symmetry-breaking, period-doublings, universality, etc., are widespread in nature: from chemical systems of autocatalytic reactions (e.g., the 'Brussellator', cellular metabolism) to ecosystems (population dynamics, as described by the logistic equation or the Lotka-Volterra equations) to economic and financial systems (e.g., mechanisms relating inflation and unemployment, represented by the 'modified Phillips curve', a particular bifurcation diagram). For many of these systems, self-organisation does not rely on fine-tuning but occurs spontaneously. In a sense, these systems adapt on the fly in response to changes in the environment.

1.3.4 Adaptivity

As shown, microscopic perturbations of the initial conditions are enough to produce unpredictable effects. In real systems, extreme sensitivity to initial conditions can obtain not only due to temporal parameter tuning but also to *spatial inhomogeneities* in the distribution of the components (reactants in chemical and biochemical systems, bits of information in socioeconomic systems, etc.), which result in a space-dependent kind of self-organisation, typical, for instance, of chemical waves and arguably responsible for, e.g., the morphogenesis of the embryo (Prigogine and Stengers, 1984, pp. 148-152).[8]

[8] Modelling this behaviour makes it necessary to represent spatio-temporal configuration of components along with variations in aggregate values of variables. For this purpose, modelling techniques different from differential equations are used, e.g., cellular automata and agent-based models (see chapter 3).

Recall the non-ideal pendulum example. Its behaviour is relatively simple insofar as it stays in one attractor basin. Complex behaviour arises when the number of agents increases and their interaction produces a more complex phase space (e.g., coupled pendula). Yet, even if the system fluctuates between distinct attractors in a unregulated way, not in response to tuning of control parameters to precise values, there can still be simple features that emerge out of the underlying complexity. For instance, if one plots the scale of some complex phenomenon x (e.g., the size of fjords, the number of biological extinctions, the magnitude of earthquakes) against the frequency $F(x)$ with which it occurs, one sees that this relationship obeys a power law distribution (see Bak, 1997, chap. 3):

$$F(x) = x^{-\alpha} \tag{1.3}$$

By taking the logarithm on both sides of the equation,

$$\log F(x) = -\alpha \log x \tag{1.4}$$

and then plotting $\log F(x)$ versus $\log x$ one obtains a straight line, α being the slope of the line. This means that many complex phenomena are *scale-free*, i.e., they have no characteristic size. Those whose power law distribution has the same exponent are said to belong to the same universality class. The mechanism responsible for this phenomenon is sometimes called 'self-organised criticality' and the critical state in which the system organises itself is regarded as an attractor (see Bak, 1997, chap. 2). According to Bak's theory, critical events are unpredictable.

An alternative interpretation of self-organisation is based on the notion of 'intermittent criticality' (Sornette, 2002). Here, the system's organisation does not depend on a permanent state of chaos, but on an 'on-off intermittency' of chaotic and non-chaotic periods, due to the heterogeneity of the system's components. The idea originates in the study of earthquakes, and has been applied to phenomena as diverse as epileptic seizures, child birth and stock market crashes. In short, heterogeneous ('disordered') materials experience more and more cracks until the critical point is reached when the main fracture is formed. The threshold to criticality may be calculated by studying the process of variation of physical quantities (acoustic emission, elastic, transport, and electric properties). Near global failure, the cumulative elastic energy released follows power law behaviour corrected for presence

of log periodic modulations, such that the exponent α of the power law is a complex number. This entails, among other things, that α is *non*-universal, but a function of the damage law. Contrary to Bak's theory, this allows for prediction by means of the so-called 'time-to-failure' analysis, based on the detection of the acceleration of some signal on the approach to global failure.

Under this interpretation, the critical point is reached not in times of chaos, but of order, when a higher-than-linear positive feedback is not compensated by an adequate negative feedback, ensuing in instability. Exogenous shocks are only the trigger of the fracture, the ultimate cause being the endogenous process by which the system gets more and more ordered.

Another way in which the inhomogeneity of the system's components is responsible for self-organisation depends not so much on spatio-temporal inhomogeneities between similar components but on the *diversity* of the components (e.g., genes or proteins in a cell, species of plants and animals in an ecosystem, investors in the market). These components, or agents, are highly autonomous, in the sense that they follow their own set of rules. However, they are also highly connected and interactive with one another—from which surprising behaviour can arise. Furthermore, agents are complex also in the sense that they can *adapt* to changes—which renders the system itself adaptive. Self-organisation due to adaptation depends on 'constructive' interactions with the environment, rather than on some blueprint in the system's components.

Ultimately, the source of this kind of complexity is the *radical openness* of the system. 'Radical openness' is a cousin of 'holism', the thesis that everything interacts with everything else. Real, open systems do not just wait to reach equilibrium within some pre-established phase space. They also actively respond to outer interactions and exchanges with the environment in which they are embedded. The action of the environment, which in conditions of closure or quasi-closure is responsible for driving the system away from equilibrium, breaking symmetries, etc., can also result in a modification of the predicted phase space, with the formation of new attractors, each attractor corresponding to a different stable configuration in a 'fitness landscape' (see Heylighen, 2001). Along this process, the system tends to change its boundaries (enlarge or shrink) and adapt so as to find the fittest configuration.

Some go as far as suggesting that radical openness makes the locution 'complex systems' an oxymoron—complex *systems* are only in the eye of the beholder (see Chu, 2011). Others, instead, only conclude that the identity of

complex systems tend to be very fragile:

> In complex systems, both the definition of entities and of interactions among them can be modified by evolution. Not only each state of a system but also the very definition of the system as modelized is generally unstable (Prigogine and Stengers, 1984, p. 204).

The phenomenon of radical openness can be illustrated with reference to what happened to an ecological system, Lake Victoria, after the introduction of an alien predator species, the Nile perch (see Chu et al., 2003). Lake Victoria used to contain more than 300 different species of cichlid fish, which comprised about 80% of the biomass of the lake. The Nile perch was introduced because more suitable to commercial fishing and export trade, with the hope that the local population could benefit from this. Ecologists predicted that cichlid fish would be driven to extinction, which in turn would leave perch without food and cause its own disappearance, with a negative, rather than positive, impact on the fishing activity. The first consequence occurred as predicted. However, what ecologists did not predict was the increase in number of other species, like the prawn and the *dagaa*, as a result of the disappearance of the cichlid fish, and the adaptive response of the perch, which modified its diet so as to incorporate other fish in the light of this sudden change. This resulted in the settling of the system on a new, unforeseen attractor. Unexpected were also other, undesired effects on the local economy of the lake area. For instance, the introduction of the perch determined an overall increase in the fish in the lake. However, given that the perch is less affordable than cichlids for the locals, it is exported rather than sold on local markets. The effect of this was a change in the population's diet, which in the future is likely to affect public health and act back on the economy of the area.

Given the inaccessibility of these phenomena to traditional analytical approaches, other modelling techniques are employed, for instance, agent-based models. However, no available technique can predict the whole network of interrelated effects that result from the radical openness of the system.

Conclusion

Although no general definition of complexity is available, it is possible to identify core features of complex systems that are responsible for phenom-

ena that are regarded as complex, for some reason or other. Among these features are nonlinearities, extreme sensitivity to initial conditions, bifurcations and symmetry-breakings, adaptivity. Complex behaviour poses serious limits to our ability to predict, control and explain. This, in turn, is meant to constitute a substantial difference from the kind of phenomena studied by classical, Newtonian, science, and seems to conjure against the meaningfulness of causal talk in complex systems. In chapter 2, I argue that complexity does not make the notion of causality dispensable, but rather prompts a revision in our notion of causality. This justifies the project of investigating the meaning of causal claims in complex systems.

Chapter 2

Complex Yet Causal

Sometimes, complexity and causality are presented as incompatible, for several reasons, e.g., chaos (in its informal associations with randomness), emergence (as a gap in a causal chain), etc. In the present chapter, with reference to the characterisation of complexity given in chapter 1, I will dispel the doubt that causality and complexity might be mutually incompatible. They are not. My argument will proceed as follows: in §2.1.1 I show that complexity does not entail indeterminism, hence absence of causality; in §2.1.2 I argue that complexity does not make the notion of causality methodologically dispensable; in §2.1.3 I argue against the view that causal facts in complex systems are 'derivative' with respect to 'fundamental', non-causal facts; in §2.2.1 and §2.2.2 I reject the incompatibility between causality and emergent behaviour in complex systems.

2.1 Are causality and complexity incompatible?

Complex systems scientists *look for* causal relations, often explicitly so. This is true both of systems biology (see Yoo et al. (2002); Blair et al. (2012); cf. O'Malley and Dupré (2005)) and of computational economics (see Chen and Hsiao (2010); cf. Tesfatsion (2006)).

Causal relations are often acknowledged as the target of complex systems scientists' practice of model building. One prominent advocate of this idea is Rosen (1985, 1991, 1998). Rosen theorised that the process of model building in science is best captured by what he calls the "modelling relation" (see figure 2.1), a relation of *encoding* of causal structures into formal structures to derive certain conclusions, and of *decoding* of the results drawn by the model with reference to the causal system one started off with. Following Rosen, the idea that models of complex systems are essentially *causal* models has become widespread among complex systems scientists (cf. Casti (1997,

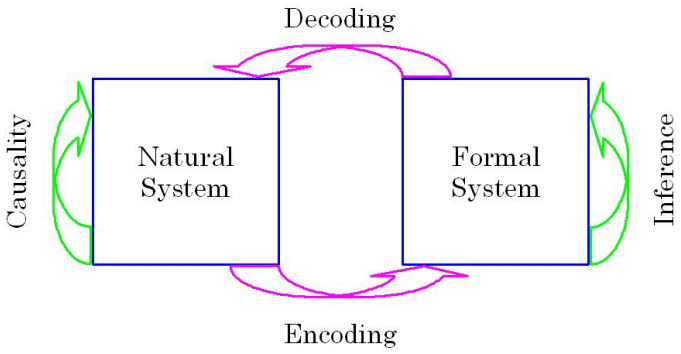

Figure 2.1: The modelling relation. Redrawn from (Rosen, 1985, p. 74).

p. 205); Mikulecky (2001, 2007); Louie (2010); Kineman (2011); Wolkenhauer (2001); Wolkenhauer and Ullah (2007); etc.). Whence the question: What is the notion of causality at work in complex systems sciences?

Recall the features that were meant to characterise classical science as opposed to complexity science (viz. lawfulness, determination and reversibility) and the doctrine of causal determinism that such features motivated. How does the discovery of complexity affect the notion of causality? Usually, complex systems scientists claim that complexity demands that we replace one notion of causality, viz. the Newtonian one, with another, more 'systemic' notion (Goldstein, 1996; Mikulecky, 2007; Érdi, 2008). On the face of this, one might think that the presence of complex phenomena makes causal language devoid of content, or inappropriate, that the problems with the doctrine of determinism that complexity brings to light make troubles for the notion of causality, too. In this section, I consider and reject various reasons that have been, or could be, advanced for such an incompatibility.

2.1.1 Indeterminism?

Self-organisation, as shown, can be more or less spontaneous, the result of fine-tuning of some control parameter or of an adaptive process. In any case, complex behaviour arises out of a process that is regarded as both deterministic and chancy:

> Self-organization processes in far-from-equilibrium conditions correspond to a delicate interplay between chance and necessity, between fluctuations and deterministic laws (Prigogine and Stengers,

1984, p. 176).

One may be tempted to conclude that since the behaviour of far-from-equilibrium systems is unpredictable, it is also indeterministic—hence a-causal. Well, this move would be fallacious.

As I said in chapter 1, sensitivity to initial conditions and openness to outer influences, typical of complex systems like Lorenz's, Bénard's, or Lake Victoria, make it hard to predict their quantitative, long-run behaviour. Does this mean that the system is *indeterministic*? Arguably, not. The fact that the behaviour of certain systems is not analytically deducible (e.g., n-body problem) need not undermine the belief that their behaviour is also caused, and that it is fruitful to search for factors responsible for it. Also, non-deducibility is compatible with the idea that irreversibility is only impossible in practice, due to our limitations. In fact, irreversibility can in principle be accounted for in terms of the microscopic, deterministic, processes (whether known or unknown) that produce macroscopic organisation.

Consider symmetry-breaking as involved in, e.g., Bénard rolls. How does it arise? If the direction of rotation of the rolls were genuinely indeterministic, then given all possible information, it would be impossible to predict whether the rotation of each molecule at the bifurcation point would be clockwise or anti-clockwise. Admittedly, the model doesn't tell us this. It deterministically predicts that rolls will emerge, that successive period-doubling will occur at a given rate and other macro-features, but it doesn't specify the direction of the rotation at each bifurcation. This is un-determined. But lack of determination or determinability must not be confused with indeterminism (see Atmanspacher (2002), Auyang (1998, pp. 266-267)). The first is an *epistemic* notion. It depends on our limited ability to pin down initial conditions exactly, to specify all interactions with the environment, etc. The second notion, instead, refers to the way reality is in itself. If reality is indeterministic, then Nature itself doesn't 'know' how the system will behave from one moment to the next. This is intrinsically, *ontically* unspecified. Bénard rolls behave indeterministically only in the first sense. Additional knowledge, in fact, does in principle contribute to the prediction of the direction of rotation. It is now believed that external fields, such as the gravitational field, insignificant at equilibrium, can affect far-from-equilibrium systems so as to determine pattern selection (see Prigogine and Stengers, 1984, pp. 163-165). Obviously, to gain predictive, or retrodictive, ability with regard to the behaviour of individual molecules, one would need to know how gravitation contributes to

affecting the molecules' position and velocity as temperature increases, which is impossible in practice. Furthermore, even if this were known, the system would plausibly remain extremely sensitive to other fields, vibrations and all sorts of fluctuations. However, evidence that symmetry-breaking can be so influenced by external factors seems to reinforce not weaken the belief that the behaviour of the system is *determined*, although the knowledge required to *determine* this behaviour would be beyond human reach.

The same reasoning applies to adaptive systems, e.g., the lake Victoria. The open character of these systems makes it even less relevant to attempt to completely eliminate the source of error deriving from misspecification of initial conditions. In fact, the system is also continuously involved in responding to changes in the environment. An external influence (e.g., a new player) can not just force the system to reach a new equilibrium but also modify the fitness landscape. Even though these changes are, in a sense, totally determined, they are often unpredictable. In fact, we are bound to represent systems as closed or quasi-closed (open to a given number of interactions). So, even if the system's response to the environment were in fact deterministic, this response would remain non-determined by the model. This does, in a way, transcend our representational ability, but does not entail indeterminism.

Complexity (in most cases, at least) need not entail indeterminism. Complex phenomena may well be the result of deterministic processes. How and to what extent, then, should the apparently deterministic and yet unpredictable character of complex systems affect our notion of causality?

2.1.2 Methodological dispensability?

Reflections on the identification between causality and determination has prompted a strand of eliminativist positions on causality, most famously Russell (1913) and, more recently, Norton (2003).[9] Russell (1913) argued that causality is "a relic of a bygone age". In short, Russell's argument can be put in the form of a dilemma: either causality is interpreted as determination, or necessitation ('same cause, same effect'), or it is interpreted as regular association ('like cause, like effect'); in the former case causality is either inapplicable or incoherent; in the latter case it is parasitic on and always

[9]The doctrine of determinism is called into doubt by our best account on the behaviour of microscopic particles, that is, quantum mechanics. However, in the following, I will proceed without relying on the assumption that reality is fundamentally indeterministic—there are independent reasons why the identification of causation with determination cannot provide a satisfying analysis of causality in complex systems.

displaced by scientific laws; either way, the principle of causality in science is dispensable. The eliminativist claims that science can do away with the *principle* of causality, which states that for any event there's a cause. One may agree on this but maintain—as I do—that the *notion* of causality is not dispensable. The eliminativist may then claim that it is the incoherence of such a notion rather than its universal applicability that demands elimination. This objection, too, I will argue, misses the target.

It is worth considering Russell's argument in more detail, because attempts to deny the meaningfulness and/or the indispensability of causal claims in complex systems are ultimately rooted in one or the other horn of Russell's argument. The first horn is related to a *metaphysical* problem, i.e., the problem of identifying causal relations with global, inaccessible, determination relations. As shown in chapter 1, complexity makes this task extremely difficult, if not impossible in principle. The upshot of the argument is that if causality means determination, the notion is either in principle inapplicable or incoherent—in either case, causality is dispensable. Since I agree with Russell here, I am not going to take issue with him on this. Rather, I will consider a generalised version of the first horn, which one finds expressed by current eliminativists, such as Norton (2003). Whereas Russell's version is about the incompatibility between the principle of causality and determinism, the generalised version is about the incompatibility between the principle of causality and the 'ultimate structure of reality'—*whatever that structure is*. The second horn, instead, has to do with a *methodological* problem, the problem of characterising causal relations in a way that renders them both accessible and non-eliminable. For reasons of exposition, I will consider the second horn in the present section and leave the discussion of the generalised version of the first horn to §2.1.3.

Whilst acknowledging that there are many regular sequences in nature, Russell denies that the aim of science is to discover such sequences (see Russell, 1913, p. 8). Russell assumes here that the only sensible methodological interpretation of 'like cause, like effect' is in terms of regular sequences of events which occur with some time-interval between them (Russell, 1913, p. 4). The existence of a time interval between cause and effect is necessary to avoid incoherence—even if this runs against the intuition that the efficacy of a cause depends on its being contiguous with the effect for the possibility of possible interferences to be blocked. In fact, so Russell reasons, if we demand (temporal) contiguity we incur a dilemma. Either causes are states involving

change within themselves, or they aren't. If they admit changes, presumably only the causes' later parts will be relevant to their effects. But since time is dense (i.e., infinitely divisible) a regress arises: for any putative causal state, there will be another state within it, viz. its later part, which is contiguous to the effect. Instead, if causes are static, we would be forced to accept a strange fact: something static determining a change. What renders the cause efficacious, and why does the cause become efficacious at some time rather than an earlier or later time, or at no time at all? Thus, Russell concludes, cause and effect must be temporally separated. However, against this notion of causality, Russell maintains that the task of scientific laws is to describe 'differential' relations, which are functional relations between continuously variable quantities, not regularities between disjoint events or states:

> There is no question of repetitions, of the "same" cause producing the "same" effect; it is not in any sameness of causes and effects that the constancy of scientific laws consists, but in sameness of relations. And even "sameness of relations" is too simple a phrase; "sameness of differential equations" is the only correct phrase (Russell, 1913, p. 14).

If causality means nothing more than regularity, then proper scientific laws will make causal claims methodologically dispensable, in the same way the laws of astronomy displace claims such as 'The night is the cause of the day'.

Along lines very similar to Russell's, Wagner (1999) has argued that the notion of causality cannot be meaningfully applied to complex systems, due to the presence of nonlinear interactions and the lack of strong regularities. Wagner's argument, too, starts with the assumption that the only meaningful notion of cause is not that of *total*, or determining, cause but of *regular* cause. In fact, if identification of total causes were necessary, this would prevent the notion from being useful, as the conditions for a state to cause another would be too unrepeatable. Instead, causes should be interpreted as states that are regularly followed by effects, when other factors in the background are stable enough or controlled for. But this is very rarely the case in complex systems, which, due to nonlinearities, are such that changes in a variable may result in changes in attractor basin, and consequently in effect. Causal talk is useful only in the restricted case of linear systems at equilibrium, where the regularity interpretation is applicable, or in the case of 'young' sciences, where only superficial knowledge of the systems studied is available.

The efficacy of this second horn of the dilemma depends, more or less explicitly, on either the *conceptual incoherence* or the *methodological dispensability* of a causal interpretation (not in terms of regularities) of scientific laws.[10]

First, the conceptual incoherence of a causal interpretation of laws seems to rest, for Russell, on the following premisses: causes and effects are crude descriptions of discrete events, whereas laws specify functional relations between continuously variable quantities; and to analyse causes as temporally contiguous to their effects—which is what it would take for causes and effects to acquire the status of the states described by scientific laws—leads to a conceptually incoherent notion of causality. However, Russell's argument for conceptual incoherence is far from conclusive. Chakravartty (2007, chap. 5), for instance, agrees that if time is dense and quantities can vary continuously in it—which is what scientific laws presuppose—then picking out a value of a quantity, that is a particular event, as the cause of another, is an arbitrary, at most pragmatically convenient choice. There is nothing intrinsic (or 'substantial', in the sense defined in §1.2) to the chosen states that makes them be 'the' cause and 'the' effect. Yet, Chakravartty puts the blame not on the incoherence of the notion of causality itself but on the interpretation of cause and effect as events rather than properties. Once we take properties as relata, we may—coherently—understand causal talk as referring more loosely to a *multitude* of *continuous* causal relations between properties.[11] Depending on the context, only some of such properties—the *relevant* ones—and only some states—'the' cause and 'the' effect—will be mentioned as salient. As a result, we get ordinary causal claims. They are not incoherent, provided one bears in mind that they are just a shorthand for more complicated, *causal* stories. This move allows one to reinterpret the relation between causal claims and scientific laws: causal claims are not 'displaced' by (non-causal) laws; rather, laws themselves may be causally interpreted; and ordinary causal claims provide legitimate stories, only more coarse-grained than laws.

This brings me to the second point, viz. the methodological dispensability of a causal interpretation of scientific laws. This depends on the implicit assumption that the discovery, interpretation and use of such laws does not

[10] Clearly, the second horn depends also on the (non-)viability of regularity accounts of causality, the criticism of which I postpone to §4.1.

[11] Chakravartty sees this as an argument in favour not just of the coherence of the notion of causality, but also of causal *dispositionalism*. For the purpose of the present argument, I need not discuss the details of Chakravartty's position—which I leave to §5.5.2.

require the notion of causality, which is debatable. For one thing, contrary to Russell and Wagner, it does not seem that causality is made dispensable by a deeper knowledge of the systems studied, as evidenced by complex systems sciences, where a lot of knowledge is available and yet causal talk is widespread (e.g.: 'an increase in greenhouse gases will cause a change in the climate'; 'temperature raising causes convection'; 'the introduction of the perch drove cichlid fish to extinction'). Admittedly, this argument may not be strong enough—these are, after all, 'young' sciences. But there is more: if laws of 'mature' sciences such as physics are symmetrical, as Russell acknowledges, then how can one select future solutions rather than past solutions for inference and explanation? Scientists need a way to perform such a selection. It seems plausible that physicists often apply (at least implicitly) a general notion of causality, that dictates that causes (e.g., the existence of a star) precede their effects (e.g., its distant correlated observations) (Frisch, 2012). This allows them to draw inferences about, and explain, such observations. If it is true that the asymmetry between cause and effect cannot be reduced to the laws (which are symmetric) plus some non-causal set of conditions, the notion of causality is *not* methodologically dispensable and is here to stay. In the words of complex systems scientist Auyang:

> Causation is not a relic but thrives in scientific reason. When formulating theories, scientists face open situations with a multitude of entangled events and processes. They must weigh various factors and judge what to include in their models, thus engaging in cause-effect analysis (Auyang, 1998, p. 259).
>
> (...) We admit mysteries and brace for catastrophes, but they are recognizable as such only against a more or less stable background of causal relations, without which the world would signify less than sound and fury (*ibid.*, p. 268).

The notion of causality is a *useful* heuristic principle and a principle of explanation to which scientists regularly appeal in practice. The second horn doesn't make causality dispensable.

2.1.3 A 'derivative' reality?

The generalised first horn of Russell's dilemma has it that the legitimacy of causality depends on the principle of causality being 'fundamentally true' to

some 'ultimate structure of reality'. The idea is that in order for causality to be a legitimate, scientific notion, there must be a formulation of the principle which is both factual and 'fundamentally true'—meaning: true under the description of mature sciences, or true under a description whose terms refer to entities belonging to the fundamental scientific ontology.[12] To put the eliminativist claim in the form of an analogy, "seeking causation in nature is akin to seeking images in the clouds" (Norton, 2003, p. 32). The analogy is clear. The image has no fundamental reality: although grounded in the shape of the cloud, the reality of the image is not backed by any fundamental property of the cloud, but just depends on us and our interpretation of the cloud. But what counts as 'fundamental' and what doesn't?

What the eliminativist seems to assume when he argues that "causality" does not refer or refers at most to facts which have a 'derivative' reality is that causal claims *misrepresent*, or are incompatible with some 'literally true' story about that reality. The issue is clearly reminiscent of the realism vs anti-realism debate in philosophy of science as construed after van Fraassen (1980), namely as a debate on which theoretical entities should be assumed in our ontology in order to make sense of what science says. In the present context, however, the issue is not on whether causality is, or should be interpreted as, a theoretical entity. Despite appearances, (almost) all parties agree on this. Rather, the issue is on whether or not causal claims—whose relata may or may not be theoretical entities—are true only to the extent that the *relation(s)* referred to by them can be redescribed literally, that is, in terms of physical laws. Roughly put, physical laws are statements describing the behaviour of fundamental, theoretical entities. Some add to this characterisation the requirement that a law describe a behaviour that holds universally; others add the requirement that the behaviour hold necessarily; and some add both requirements. So, at bottom, the issue that causal claims are not fundamentally true is that they are not underpinned by *universal* and/or *necessary* laws.

On the one hand, it is true that complex systems sciences show that complex phenomena are not subsumable under non-exceptionless laws. So, for instance, many dynamical 'laws' (e.g., kinetics in chemistry and molecular biology) are regarded not as strict laws, but only as contingent generalisations:

[12] An argument against the generalised first horn will allow me to ignore the common claim that given the *fundamentally indeterministic* nature of reality, there is no causation *anywhere*.

> In contrast with close-to-equilibrium situations, the behavior of a far-from-equilibrium system becomes highly specific. There is no longer any universally valid law from which the overall behavior of the system can be deduced. Each system is a separate case (Prigogine and Stengers, 1984, pp. 144-145).

On the other hand, however, it is not clear why this makes an argument against causality in general, rather than merely against its interpretation as necessitation.

My concern with the generalised first horn is that it trades one fundamentalism, viz. reality being governed by the principle of causality, for another, viz. there being a true story on the ultimate structure of reality, a 'theory of everything', with respect to which the meaning of any statement is analysed as referring to either a 'fundamental' or to a 'derivative' fact, and such that causal relations exist if and only if reducible to the laws of this theory.[13] This attitude is often—derogatorily—labelled "fundamentalism" by so-called 'scientific pluralists' (cf. Stewart (1997, chap. 17), Cartwright (1994), Giere (2003), Mitchell (2003, chap. 6)). I think there are good reasons to resist the fundamentalist temptation on pluralist grounds and to investigate causality in complex systems in its own right, that is, with reference to the local workings of systems of interest rather than to some fundamental reality, whether global or local, deterministic or non-deterministic. Here, I will summarise two arguments against fundamentalism, namely Cartwright (1994) and Giere (2003)'s, which I find particularly convincing.

Giere (2003) argues that claims from different disciplines on different aspects of reality are all legitimate, provided they meet some empirical standard. Clearly, when conflict arises among these claims, unification of the perspectives and solution to these conflicts is desirable. However, Giere maintains, such a unification process need not be interpreted as driven by the metaphysical belief that good science depends on the reduction of all perspectives to a theory of everything, but rather by a sort of methodological maxim, to be followed when empirical evidence suggests that it is likely to promote achievement of the goals of scientific inquiry. In Giere's words: "proceed on the assumption that there is a single world with a unique structure" (Giere,

[13] As I see it, this view could lead both to elimination and reduction. The first option is advanced by the eliminativists. The second option is advanced by those who aim to reduce informal causal talk to talk about either 'fundamental' causal relations (Salmon-Dowe account) or non-causal 'X-' states of affairs (regularities, counterfactual dependences, probabilistic dependences, etc.) (see chapters 4 and 5).

2003, p. 17), a "causal" structure (Giere, 2005, p. 158). How can this maxim be justified with reference to causal claims themselves? The best justification for taking well-confirmed causal claims as non-derivatively true is pragmatic. As I said in §2.1.3, the search for causal relations has proved very fruitful so far. If anything, this should tell us that causal talk is not be eliminated on the ground of what the ultimate structure of reality looks like.

In a similar vein, Cartwright (1994) argues that we have no reason to believe that claims—causal claims included—that we have established under certain special circumstances (controlled, ideal-experiment in fundamental physics) will *always* be universally exportable. On the contrary, evidence suggests that all laws are *ceteris paribus*, valid in certain domains not in others. Analogously, models are applicable to certain domains not to others. But luckily, we don't need universally valid claims for scientific language to be able to represent and explain. Accordingly, the scientific enterprise is best interpreted as the production of a 'patchwork of laws' rather than the attempt to build a 'theory of everything'.[14]

One consequence of holding a pluralist position is that there is no incoherence in denying then the truth of determinism whilst admitting the existence of deterministic systems (cf. Auyang, 1998, pp. 262, 267-268). All that is required for the latter is portions of reality which are well-representable, for some time and to a good degree of approximation, by means of deterministic models. And wherever such models work, talk of causal relations seems perfectly legitimate.[15] To say that such causal relations have only 'derivative' reality will only be convincing to those who believe that all sciences inherit their legitimacy from fundamental physics, whose stories are the only admitted as fundamentally true. But according to the pluralist, scientific stories are not 'fundamentally' true, not even by the scientists' own light. Rather, they are devices, and as such can be better or worse depending on the use we make of them. Models of complex systems are no exception. They are good representations of the domains where they do work, but say little about the domains where they do not. And when they get the story right, the extent to which they represent cannot be measured independently from the uses that the representation is put to.[16] So, although there may be a true story, or a

[14]For Cartwright, this argument is meant to buttress the claim that scientific success is enough to infer to the existence of causal powers, or 'capacities'. For a discussion of Cartwright's views on causality, see §4.3.1, §5.5.1, and §6.3–§6.4.

[15]This is not to say that causal talk is legitimate *only if* systems can be modelled deterministically.

[16]I'm sympathetic towards Morgan and Morrison (1999)'s intepretation of models as

good model, it makes little sense to demand that a story be 'fundamentally' true, or a model 'literally' representational.

2.2 Causality and emergence: a difficult liaison?

An objection which may be advanced to undermine the attempt to investigate the meaning of causal claims in complex systems is that emergence conflicts somehow with causality. As I am going to argue, this objection is unfounded, and largely depends on the implicit endorsement of fundamentalist assumptions.

The objection can be put as follows: if emergence were to exist (somewhere), its existence would entail absence of causality (at least somewhere). The intuition here is that emergent behaviour cannot be traced back to specific causes, hence cannot be causally explained. So, even if we don't demand that the notion of causality be universally applicable—as the advocate of the principle of causality maintains—it may not apply exactly where we would like causal talk to be meaningful. If complex systems phenomena involve emergence, this may be bad news for the project of analysing causality in complex systems. Sometimes the objection is accompanied by the corollary observation that emergent phenomena would have 'downward' causal influence on their base, and the notion of downward causation makes the notion of causality problematic: if causation runs both upwards and downwards, then the asymmetry of causality is lost. An argument is needed in defense of my project against this objection.

Although there is no agreement on what 'emergence' means, my argument does not rely on the existence of an unanimous definition. Rather, I proceed by breaking down the problem into sub-problems, each arising from a different feature of 'emergence' that may motivate skepticism about the meaningfulness of causality in complex systems. In particular, I will only consider *prima facie* emergent phenomena that obtain in complex systems, viz. macroscopic *patterns* of the system that arise out of the microscopic behaviour of its parts (plus interactions with the environment) (Humphreys, 2008a). A 'pattern' is understood as a non-random property of the system, which is distinct from any properties possessed by the initial state of the system. The behaviour that is responsible for the pattern is characterised by some updating function (e.g.,

'mediators', especially towards an *inferentialist* reading of this view. For more on how my proposal relates to the referential function of causal language, see §8.1.2.

difference equations, some suitable algorithm) that transforms each state of the system into its successive state. In what sense are such patterns emergent?

The meaning of 'emergence' has several facets (Humphreys, 2008a,b). First, one can distinguish *ontological* emergence, i.e. the appearance of some novel property, from *epistemological* emergence, i.e. the impossibility to calculate (or reduce) the emergent property from (to) microphenomena. Secondly, one can distinguish *diachronic* emergence, i.e. the emergence that results from a temporally extended process, from *synchronic* emergence, where the emergent state and its lower-level base are simultaneously instantiated.

2.2.1 Diachronic emergence

In the case of diachronic emergence (e.g., a chemical clock or an embryo arising out of microscopic, disordered interactions among molecules), it is easy to see where the alleged incompatibility with causality could lie. Both diachronic emergence and causality involve temporally extended, asymmetric processes. So, it might be that whereas a process leading up to an effect can also be traced back to its cause, this is not possible in the case of a process that terminates into an emergent phenomenon. Yet, this seems very strange. After all, diachronic emergence is the result of a process that starts somewhere. It may be true that (the selection of) a given starting point is not sufficient for (predicting) the emergent state—in which case that starting point does not qualify as a total cause of it. However, as argued in §2.1, causality need not be identified with determination. That diachronic emergence is not incompatible with causality in complex systems is best shown by uncoupling the epistemological and ontological side of the issue.

According to the *epistemological-diachronic* view (see Humphreys, 2008a, p. S584), a state or a property instance is emergent with respect to a domain D iff it is impossible, on the basis of a complete theory of D[17], to effectively predict that entity or to effectively compute a state corresponding to that feature. A recent, popular way to cash out this idea is the doctrine of *weak emergence*. A weakly-emergent property is usually defined with respect to a *model*, and need not have ontological implications as regards the features of real systems.[18] In short, the process leading up to the emergent state

[17] 'Complete' theories are theories satisfying conditions such as universality (as when D is physical domain) and closure (as when D is the domain of a well-established science).

[18] A weakly emergent property can be defined as follows. Assume P is a system's higher-level property which is in principle incapable of being possessed by the system's micro-

is computationally incompressible, so the only effective way to discover how the system will evolve is to let the computational model work out its own development. Most complex phenomena turn out to be emergent in this sense. This position, however, is not incompatible with treating emergent properties as causes or effects. We can account for weak emergence in complex systems in the phase space terminology introduced in §1.3.2. An emergent state, or pattern, can be interpreted as an attractor (Silberstein and McGeever, 1999; Emmeche et al., 2000). For instance, organisms can be regarded as consisting of highly complicated attractors for the behaviour of molecules in a biochemical space. In this interpretation, cell types (around 250) are attractors for the dynamics of thousands of genes.

The ontological view, instead, regards a property as emergent iff it is 'genuinely novel', i.e. a property with novel causal powers with respect to its emergence base. Such powers are reflected in laws which connect complex physical structures to the emergent features. Emergent laws are fundamental, i.e., irreducible to laws characterizing properties at lower levels of complexity, even given complete information on boundary conditions (O'Connor and Wong, 2009, §3.1). This view is part and parcel of the fundamentalist picture of a physical world entirely constituted by physical structures, such that the hierarchy of composite structures corresponds to distinct levels of objects of increasing complexity, each level obeying its own set of physical laws.

There are two main accounts of *ontological-diachronic* emergence. One is the 'dynamic' model of emergence developed by O'Connor and Wong (2005), the other is the 'fusion' model proposed by Humphreys (1997a,b). The former is meant to capture the emergence of conscious states (e.g., visual awareness, qualia). The latter focusses on quantum entanglement and the directly observable phenomena due to it (e.g., superconductivity and superfluidity in helium, spontaneous ferromagnetism that occurs below the Curie temperature), although one may view 'interactivist' (or 'process') ontologies as extensions of this model that target also higher-level phenomena such as the emergence of autonomous and/or intentional systems (see, e.g., Campbell, 2009). I need not go into the details of these models. It suffices to say that

level constituents, but which is structural, i.e., expressible in terms of (reducible to) such constituents and their spatial relations. Then P is weakly emergent iff P is a non-random property of the system, distinct from any property possessed by the initial state of the system, and derivable from all of the system's micro-facts but only by simulation. This is roughly Bedau (1997, 2002)'s definition, with Humphreys (2008a,b)'s addition that the emergent state be *non-random*, i.e., its description cannot be shorter than a conjunctive specification of all the constituent microstates.

both models rely on the possibility that higher-level properties of the emergent state change without changes in properties of its lower-level base, in a way that ultimately depends on the presence of some fundamentally indeterministic process, such that features of the emergence base underdetermine which among several global states is instantiated. Now, since emergence here seems to depend crucially on the presence of indeterminism, one may worry that indeterminism could propagate to higher levels in a way that renders causal talk inappropriate. But the worry is unfounded. First, it is simply false that all complex systems phenomena involve genuine indeterminism. On the contrary, as shown in chapter 1, many of them can be successfully modelled by assuming that they are deterministic. Secondly, even if indeterminism were to sneak in somewhere (e.g., Bénard rolls), this would not *ipso facto* make the emergent phenomena uncaused. Most of their determinants can be identified, so that indeterminacy is narrowed down considerably. This makes causal talk perfectly legitimate.

Notice that talk of downward causation in the diachronic case is innocuous. Downward causation (e.g., an organism controlling its own metabolism) can be uncoupled into a constitutional, synchronic, inter-level step (the organism is made of organs, cells, enzymes, etc.) and a causal, diachronic, intra-level step (some of the organism's parts do the causal work, i.e. they digest, assimilate, etc.) (cf. Craver and Bechtel, 2007). This allows for the possibility of non-vicious, 'circular' causation: the same state type (e.g., a hunger state) can both cause (more metabolic processes) and be caused (by insufficient metabolic processes), but at different times, and no token state can both cause and be caused. Notice that this applies not only to biological mechanisms but also to social mechanisms (e.g., a bull market is constituted by positive attitudes of the traders, and such that a raise in an asset's price causes traders to buy, which in turn causes the price to rise).[19]

2.2.2 Synchronic emergence

Let us now consider the case of *synchronic* emergence. Here the problem is of a different nature. In this case, 'emergent' is usually understood as 'supervenient but irreducible'.[20] A higher-level property M (e.g., a mental property)

[19]For more on social mechanisms, emergence and downward causation in the social domain, see Elster (1989, 1998); Hedstrøm and Swedberg (1998); Bunge (2004); Sawyer (2004); Little (2006). For more on mechanisms and mechanistic accounts of causality, see chapter 5.

[20]This is usually taken to have both epistemological and ontological implications.

is said to *supervene* on P iff, if S instantiates M at t, then necessarily there is some P such that S instantiates P at t, and anything instantiating P at any time instantiates M at that time. A higher-level property M is *reducible* to P if M and P have identical causal role (cf. Kim, 1999, pp. 10-12). In what sense is emergence, so defined, incompatible with causality? The worry here is that higher-level, emergent properties (e.g., the property of the pattern), are *epiphenomenal*, that is, have no causal powers, so cannot enter causal relations. Kim (1999, 2005) has advanced an argument, the 'causal exclusion argument' (CEA), that may threaten the causal efficacy of 'higher-level' properties, i.e. the properties that special sciences—complex systems sciences included—talk about.[21] If CEA is successful, causal relations in complex systems are either spurious or 'derivative'.[22] Contrary to this reading, I want to show that CEA points at most to a tension that derives from the fundamentalist's metaphysical assumptions, viz. that the ontology of reality is layered and that the objects at each level obey their own set of laws, assumptions that lurk in the background of almost all the debates over the nature of causal relations. However, a scientific pluralist (in the sense defined in §2.1.3) need not endorse such assumptions: the tension disappears, and so do the worries about the derivative nature of higher-level causal powers and relations.

In short, CEA states that if a higher-level (e.g., mental) property M 'supervenes' on a physical property P without being identical with/reducible to it, the causal work attributed to the supervenient property is already done by the subvenient property and the supervenient property is causally inefficacious. So, if we want to allow higher-level causation, we must admit reducibility. How does this translate into complex systems talk? Consider an emergent pattern (e.g.: a Bénard roll; a chaotic attractor; a 'glider' in a CA), and the microbehaviours that sustain the pattern. Take two higher-level states, for instance two patterns. If the first pattern causes the second, this is in virtue of the microconfiguration underlying the pattern doing the causal job, not in virtue of the property of 'being such-and-such a pattern'. So, the latter is reducible to the former.

[21] CEA was initially proposed to argue that mental properties' causal powers are reducible to the properties of their physical substrates. However, CEA generalises to other higher-level properties (social, biological, etc.), so that their causal powers are preempted by lower-level physical properties and 'drain away' (Block, 2003).

[22] Notice that CEA motivates a kind of eliminativism which differs from Norton's (discussed in §2.1.3). Whereas for Norton *prima facie* causal facts can be reduced to fundamentally *non*-causal facts, for Kim *prima facie* (higher-level) causal facts can be reduced to fundamental *causal* facts, in the sense of the Salmon-Dowe account (see §5.4.4).

What does reduction amount to? Two possibilities. First: M is *identical* to P and conserved as causally efficacious. Here M causes in virtue of being P. Second: M is just a designator which is multiply realisable by different physical kinds. Here M can be construed as logical disjunction of two or more physical properties, and *eliminated*. In fact, M causes in virtue of being P_1, or P_2, or... Since M does not stand for any specific property which is causally efficacious, it cannot enter proper scientific laws. Kim seems to think that CEA only affects multiply realisable properties (e.g., jade, whose instantiations are samples of either jadeite or nephrite[23]), but leaves untouched 'micro-based' properties. Micro-based properties are properties that do not apply to any of the parts' proper subparts, and confer to the whole novel causal powers, hence qualify as genuinely higher-level properties. For instance, 'being viscose', 'being dense', 'being a solvent', etc. are micro-based properties of the property 'being water', properties that apply to neither hydrogen or oxygen atoms alone nor to a H_2O molecule in isolation. Kim argues that in the case of micro-based properties, it is reduction itself that stops the causal drainage. In fact, a series of reductions can be obtained for micro-based properties such that any of them is identical to its re-description at the lower level, each property in the series being causally efficacious. For instance, 'being water' = 'having such-and-such micro-based property M_L at L' (i.e., 'being H_2O') = 'having M_{L-1} ' (at $L-1$) = 'having M_{L-2}', etc.

But this move is even more devastating. Micro-based properties are, in general, *multiply composable*, or multiply instantiable, i.e., they allow more than one decomposition, or structural instantiation (see Glennan, 2010, pp. 375-376). This means, among other things, that they have certain causal powers in virtue of instantiating different composition relations, or structures. For instance, patterns are usually multiply-composable, the precise specification of their composition being irrelevant to their identity (Humphreys, 2008a). If we follow Kim's reasoning, multiply composable properties are not genuine scientific kinds (they are causally heterogeneous) and so must be eliminated. The reduction demanded by multiple composability results in an even more fine-grained fragmentation of the higher-level kinds, such that also properties such as 'having a certain mass' are eliminated.[24] Ultimately, only

[23]The idea is that 'being jade' can be construed as the disjunction 'being jadeite OR being nephrite'. Jade inherits its causal power either from one or from the other, so it can be eliminated as a genuine kind. Any scientific generalisation is which 'jade' appears is true in virtue of either jadeite or nephrite's causal powers.

[24]Perhaps even properties such as 'being water' should be eliminated if, as some suggest, water is not identical, or reducible, to H_2O (cf. Weisberg, 2005).

a few, 'purely-natural' kinds would be left. (Wimsatt (2007, p. 287) characterises such kinds as 'totally aggregative', such that the mode of composition of their parts makes no difference to the property of the whole.) However, science commonly regards multiply-composable kinds (rigidity, temperature, being water, etc.) as causally efficacious and explains (causally) by invoking them. This makes Kim's move a last resort.

If we want to reject Kim's conclusion, we need an alternative story for the causal efficacy of higher-level properties. The standard ways to give such a story are either to deny exclusion, and claim that there is no overdetermination, or to deny closure, and claim that the bottom level is not causally closed. However, neither move seems satisfying.

In the case of overdetermination, one should distinguish cases where the putative causes overlap in space-time (a pattern and its underlying microconfiguration with regard to some later state of the system) from cases where they do not (two stones thrown against a window). Only in the latter case talk of overdetermination seems legitimate—one could have one cause without the other being present. This is not the case of supervenient properties since, by assumption, the higher-level property is such that it cannot be modified without changes in the lower-level one.

In the case of causal incompleteness, one would need to argue that any attempt to reduce the higher-level property to lower-level ones necessarily misses some important aspect, either because the identities that result from reduction are not exact (Dupré, 1993), or because considering the structure of micro-based properties (Bechtel and Richardson, 1992) or the mechanistic organisation in which such properties are embedded (Glennan, 2010) amounts to attributing causal significance to something which is a level up with respect to the 'disembodied' materials in the structure. However, failed attempts at precise reductions need not entail that exact identities are unavailable in principle. Analogously, it is not clear why structural properties or mechanisms cannot be—in principle—redescribed in the terms of some lower-level theory.

My diagnosis is that the problem requires a more radical solution. It originates not from endorsing one or the other assumption behind CEA, but rather from endorsing the more general, fundamentalist assumptions that reality is layered and exhaustively describable in terms of nomological relations. A scientific pluralist need not share these assumptions. Irrespective of whether reality is causally closed, there is no reason to expect that our perspectives on reality—our models and theories—be closed. Generalisations formulated

from within a perspective need not have universal scope. Nor is it necessary that for any event its cause be describable by using only the conceptual resources available to that perspective. Nor is there reason to believe that unification of perspectives will lead to some ultimate perspective, a sort of Nagelian 'view from nowhere'. Higher-level carvings might well be eliminable, in favourable conditions, to the advantage of more fine-grained carvings. And yet, if there is no ultimate perspective, fine- and coarse-grained carvings may be equally legitimate, provided they meet some reasonable standard of empirical adequacy.

Notice that the rejection of the fundamentalist ontology need not commit one to any deep skepticism as regards the belief that reality is—somehow—structured, only the structure is messier and more complex than we may think. Accordingly, the ontology of complex systems may be best envisaged as 'tropical' rather than layered:

> (...) a heterogeneous, tropical rainforest, with converging overlapping branches, and patterns of intersecting order, residents, and connections at a variety of levels, but no single stable foundational bedrock that anchors everything else (Wimsatt, 2007, p. 12).

In this rainforest of objects and relations, the existence of stable generalisations is a fortunate phenomenon, not a necessary fact about the whole of reality. Levels are "local maxima of regularity and predictability in the phase space of alternative modes of organization of matter" (Wimsatt, 2007, p. 206). In other words, talk of levels signals the presence of particularly fruitful classifications and generalisations. However, this presupposes neither that classifications will map onto neatly distinguishable levels, nor that generalisations will hold universally. On the contrary, models and theories may require the resources of other models and theories to account for certain phenomena—in this sense, our perspective on reality are 'open'. Levels can criss-cross each other, since properties that enter generalisations at one level may be required to explain phenomena at another level.

As a result, we get both 'upwards' and 'downwards' causation. But this does not entail any causal 'symmetry'. In fact, downward causation can be explained as the higher-level pattern's local stability and insensitivity to perturbations, which can be quite naturally taken as 'governing' the behaviour of the lower-level entities (see Emmeche et al., 2000, p. 29). Here, emergence makes the relation between lower and higher state robust. When the pattern

tends to drain many different initial conditions (e.g., in systems that develop towards equilibrium), it is then quite natural to say that it 'subsumes' those states—it is a *type* of which the single phase-space points are *tokens*. However, in other cases (e.g., systems in a chaotic regime o an intermittency of chaotic/non-chaotic regimes) emergence makes the relation fragile. In such cases, talk of 'downward' causation is intuitively inappropriate.

In sum, causal claims may legitimately relate kinds belonging to different levels, provided the inferences to whose correctness they contribute are *robust* (see chapter 8). Causal claims relating kinds belonging to the same level tend to be, but need not be, the maximally robust ones.

Conclusion

I have examined the compatibility of complexity and causality. Complexity reveals the inadequacy of the Newtonian notion of causality as determination. Yet, I argued, causality and complexity need not be incompatible. Complexity does not entail indeterminism, only indeterminacy. The notion of causality is neither methodologically dispensable nor a folk science notion, that applies to facts with a 'derivative' nature. Nor is there an incompatibility between causality and emergence, whether this is interpreted diachronically or synchronically, epistemologically or ontologically. In particular, the ideas that causal phenomena have derivative reality and that higher-level properties are not genuine scientific kinds follow from fundamentalist assumptions, which one need not endorse. This legitimates the project of investigating the meaning of causal claims in complex systems.

Chapter 3

Modelling Complex Systems

In this chapter I describe the two case studies I'll use in the remainder of my thesis to buttress my argument: one from systems biology, viz. *apoptosis* (the mechanism by which cells whose DNA is too damaged commit suicide); the other from computational economics, viz. *asset pricing* (the mechanism by which the value of an asset—in finance, e.g., a stock—is determined). I introduce the *modus operandi* of systems biology and computational economics in §3.1.1 and §3.2.1. In §3.1.2 and §3.2.2 I describe the mechanisms for, respectively, apoptosis and asset pricing. In §3.1.3 and §3.2.3 I present models developed to account for these phenomena. Finally, in §3.3 I argue that the novel methodologies of systems biology and computational economics are compatible with the causal interpretation of such models. This triggers the question I'll be concerned with in the following chapters, viz. 'What interpretation of causality suits best the causal claims arrived at with the aid of these models?'[25]

3.1 Modelling biological complexity

3.1.1 Systems biology

Living organisms have developed functions that enable them to survive and reproduce. The goal of systems biology is answering questions such as 'How do these properties emerge from the interactions between the molecules that make up cells and how are they shaped by evolutionary competition with other cells?' (Hartwell et al., 1999, p. C47). The novelty of systems biology lies in the attempt to answer these questions by means of a different methodology, consisting (among other things) in numerical computations and simulations of biological processes.

[25] §3.1.1 and §3.1.2 are based on (Casini et al., 2011, §1 and §3).

The underlying idea is that experimental procedures alone, based on decomposition of the system into parts, their classification and the study of their role by knock-down interventions are not sufficient for understanding the coming about of the phenomenon. For instance, systems biology opposes simplistic generalisations like 'one gene one phenotype', or 'one protein one function'. The contribution of no single part is (usually) sufficient to bring about complex phenomena that take place in, e.g., a cell, the nervous system, or a tumour. Understanding and control of complex biological systems requires an understanding of how parts operate *together* (Lazebnik, 2002).

The sheer complexity of the biology of a cell is hardly tractable, if not overwhelming. This is due both to the chemical and physical details of each individual signaling pathway and to the vast number of interactions taking place. Different modelling techniques are used to cope with different sources of complexity.[26] Let me give a taste of the details one may want to consider.

First, one can start with the simplifying assumption that reactants are homogeneously distributed in the reaction space and that the number of collisions between molecules, by means of which chemical reactions take place, is proportional to the product of the concentrations of the reactants. The kinetics of the reactions can be described by means of ODEs. When a product feeds back into the reaction and modifies its rate, as in autocatalytic processes, or several reactions take place together so that more pathways interact, thereby forming a network, complex behaviour can obtain. Properties responsible for a variety of effects in need of explanation can 'emerge'—in the sense that they are possessed only by the whole not by the individual pathways. The study of these networks is of primary concern for systems biologists (Bhalla and Iyengar, 1999), who, often using parameter values derived from experimental studies, combine several individual pathways and test the resulting models against available data. This methodology can explain how combinations of positive and negative feedback can result in, e.g., bistability between steady states (that is, sufficiently stable equilibria), well-defined input thresholds for transition between states, prolonged signal output, etc.[27]

But there are further levels of complexity. We have so far assumed that the reactants are homogeneously distributed. This, however, is very often *not* the

[26] For reviews of modelling techniques used in systems biology and their applications, see Bhalla (2003); Materi and Wishart (2007); Bedau (2003).

[27] Bhalla and Iyengar (1999) illustrate this methodology by modelling the signalling networks which contribute to LTP (long-term potentiation). In §3.1.3, I illustrate how these procedures are applied to the study of apoptosis.

case. Fluctuations due to dishomogeneities and compartmentalisation of the components in distinct areas of the cell can often play a relevant role (Weng et al., 1999; Bhalla, 2003). Small fluctuations always lead to differences in concentrations and thus to diffusion, which introduces a kind of instability which is not just temporal but also *spatial*. The dynamics of these spatial asymmetries may be described by using reaction-diffusion equations, that is, partial differential equations where the behaviour of the variables depend not just on time but also on space. However, when the dynamics require description in arbitrary three dimensional geometries (not confined to, e.g., spheres, cubes or cylinders), then reaction-diffusion equations become very difficult to solve. This makes it necessary to partition the space into fine grids and apply finite difference or finite element methods. Also, given the non-homogeneous way in which reactions take place, often stochastic methods must be employed.

Boolean (or logical) networks (LNs) are used to reproduce the qualitative behaviour of large networks. A LN consists of a finite number of variables assuming discrete values (e.g., ON/OFF), the states of which at each time step are governed by some logical, or Boolean (AND/OR, etc.), function of the states of the variables that provide input to them. In their synchronous, deterministic, variant, LNs have the property of reaching attractors, either fixed points or cycles. Their main limitation is that their discrete time steps dynamics does not tell whether these properties are indeed biologically relevant or take place on insignificant timescales. Also, not all simulated trajectories can actually obtain in real cellular contexts. However, the qualitative study of LNs (their attractors and basins) can still highlight key design principles of biological networks. For a detailed example, see §3.1.3.

Another important class of models are cellular automata (CA), which can be used to model both temporal and spatiotemporal processes using discrete time and/or spatial steps. CA consist of large numbers of nearly identical components interacting with one another on a grid. Their states evolve synchronously by following a set of rules according to which the state of each site is determined by the previous states of the neighbouring sites. Agent-based models (ABMs) are a variation on the CA model in which objects are heterogeneous and capable of motion. ABMs will be described in more detail and illustrated with reference to the mechanism of asset pricing in §3.2.3.

The complexity of biological systems will be illustrated with reference to the phenomenon of cancer, and in particular, of apoptosis.

Cancer is a complex phenomenon, initiated by exposure to DNA-damaging factors and leading, through a succession of steps, to "unregulated cell growth" (King, 2000, p. 1). When the DNA is damaged, a well-functioning cell reacts via defense mechanisms known as 'DNA repair mechanisms', which heal the cell after damage has arrested its cycle. Depending on the kind of damage (e.g., single-strand, double-strand, mismatch), different enzymes are recruited to fix the damage. As a last resort, if the damage is serious and cannot be effectively repaired, the cell either enters an irreversible state of dormancy ("senescence") or commits suicide ("apoptosis"). However, when none of the above strategies is effective, damaged cells keep growing and dividing and, in so doing, produce mutations, which are the first step towards cancer development. Notice that, although DNA replication mechanisms are very precise, DNA damage due to both internal and external factors can produce a daily number of lesions high enough to be dangerous (King, 2000, p. 125). This is why the mechanisms that allow the cell to correct errors before they are replicated (repair) or prevent mutations (senescence and apoptosis) are so important. In fact, a mutation can start a cascade of mutations, because of its capacity to impair the cell's activities (among which there is the production of enzymes needed in DNA repair itself), so that further mutations occur more easily.

Due to its astounding complexity, cancer has been dubbed a "systems biology disease" (Hornberg et al., 2006), which needs tackling by means of a systemic approach, involving the integration of diverse evidence (e.g., scientific and clinical measurements across the entire biological scale, from molecular components to systems, both *in vitro* and *in vivo*, and from the genome to the whole patient) in mathematical and computational models to be used for generation and confirmation of hypotheses, explanation, diagnostic and prognostic prediction, and treatment (see §3.1.3).

3.1.2 Apoptosis

Apoptosis is one of the mechanisms on which the cell relies to oppose DNA damage.[28] Apoptosis depends on an intracellular proteolytic cascade mediated by caspases, a family of proteases with a cystein at their active site, which cleave their target proteins at specific aspartic acids. Caspases are

[28] The following summary develops discussion in (Weinberg, 2007, chap. 9), and the reader is referred to this text for a detailed description. A shorter yet clear introduction can be found in (Klipp et al., 2009, pp. 132-135).

synthesised as inactive precursors, or procaspases, and become activated by proteolytic cleavage. Caspases involved in apoptosis are classified as 'initiators' (caspases 2, 8, 9, 10) and 'executioners' (caspases 3, 6, 7). Caspase cascades can be triggered in several ways. The literature distinguishes between the extracellular, or extrinsic, apoptotic pathway and the intracellular, or intrinsic, apoptotic pathway. The apoptotic signal can also be amplified via a crosstalk between these pathways.

DNA damage is measured by using expression levels of DNA damage response genes, e.g., *p53*. After noticing the presence of metabolic disorder or genetic damage, protein p53 can induce cell-cycle arrest, activate DNA repair proteins (e.g., DNA polymerases), or—if damage is too severe to be cured—lead to cell death (i.e., apoptosis).[29] In a well-functioning cell, 'wild' (e.g., non-mutant) p53 normally goes through a rapid degradation, due to its being 'tagged' by the Mdm2 (murine double minute) protein and subsequently 'digested' by proteasomes. The amount of p53 increases when, e.g., its phosphorylation due to genotoxic factors or the phosphorylation of Mdm2 results in Mdm2 being unable to bind to p53. Interestingly enough, p53 promotes synthesis of Mdm2, thereby contributing to its own inhibition in a negative feedback loop. This loop successfully regulates apoptosis unless the gene *p53* mutates. In the latter case, mutation of *p53* prevents Mdm2 from binding to p53 and, as per the wild case, this results in an increase of p53. However, the defective p53 has lost its ability to act as a transcription factor, that is, is unable to bind to the promoters of genes that synthesise proapoptotic proteins in the successive stages of the mechanism.

According to available data, gene *p53* is mutated in 30% to 50% of commonly occurring human cancers (Weinberg, 2007, p. 310). The crucial, *causal* role of the protein p53 is explicitly recognised, as is the possibility of building a mechanistic model around *p53* to explain how alarm signals stop the cell cycle or trigger apoptosis (Weinberg, 2007, p. 316-317).[30] In the well-functioning cell, increased p53 takes part into both the intrinsic and the extrinsic apoptotic pathways, which I now turn to describe.

Intrinsically, p53 acts as transcription factor for the encoding of proapoptotic proteins that, by opening the mitochondrial membrane channel, allow release of cytochrome *c*. Proapoptotic proteins belong, together with anti-

[29] Following the lead of the biological literature, I will use here the same name to refer to a gene and the protein it codes for, and distinguish the former from the latter by italicising it (e.g., *p53* stands for the gene, p53 for the protein).

[30] For an example of one such model, see Casini et al. (2010, 2011).

apoptotic proteins, to a family of proteins named the "Bcl-2 family" (after B-cell lymphoma 2, the first protein found to contribute to regulation of apoptosis besides p53), due to their sharing a common coding sequence. Their balance determines the opening of mitochondrial membrane and the release in the cytosol of cytochrome *c*, that binds a protein called Apaf1 (apoptotic protease activating factor 1) and activates procaspase 9. Caspase 9, in turn, initiates a cascade of caspases 3, 6 and 7 that results in the disintegration of the cell. Executioners can be inhibited by IAPs (inhibitors of apoptosis) proteins, which in turn can be inhibited by another protein, Smac (second mitochondria-derived activator of caspases), also released by the mitochondrion together with cytochrome *c*. Levels of apoptosis are measured by using expression levels of caspases 3 and 9 as surrogates.

Let us turn to the extrinsic pathway. This is due to ligands in the extracellular space (e.g., FasL) belonging to the TNF (tumor necrosis factor) protein family, that bind to death receptors on the surface of the cell (e.g., FasR). This, by prior activation of initiator caspases 8 and 10, triggers another cascade of caspases 3, 6 and 7; p53 contributes to this process by promoting the expression of the genes encoding the Fas receptor, thereby increasing the cell's responsiveness to extracellular death ligands, specifically FasL. The extracellular apoptotic signal can be amplified by crosstalk between the two pathways: caspases 8 and 10 cleave Bid (BH3 interacting domain death agonist), which acts as proapoptotic protein that inhibits Bcl-2 antiapoptotic proteins.

Cancer cells inactivate apoptosis in several ways that enable them to survive and thrive. They can increase the level of antiapoptotic proteins, change the gene coding for p53 or its upstream regulators, methylate promoters of proapoptotic genes, interfere with the release of cytochrome *c*, inhibit caspases via overexpression of IAPs, etc. On the other hand, overexpression of proapoptotic proteins or dysfunction of antiapoptotic proteins due to mutations can result into too much apoptosis and cause other pathological conditions, e.g., Alzheimer's or Parkinson's disease.

3.1.3 Modelling apoptosis

Models of apoptosis are used for, e.g., identifying key factors responsible for typical characteristics of the apoptotic signalling cascade, such as *bistability* (eventually, the cell is either dead or alive) and *irreversibility* (once initiated,

apoptosis leads irreversibly to cell death). This, in turn, can provide useful indications for the choice of therapeutic targets and development of drugs. The choice between more quantitative and more qualitative models depends on a trade-off between computational capabilities and timescales on the one hand, and level of detail on the other. Usually, ODEs are good for precise quantitative modelling but current technology can only deal with a limited number of equations, and thus either small portions of the apoptotic machinery or very rough representations of larger portions. At the other end of the scale, LNs can comprise a much larger number of variables, hence of pathways involved. However, the results they provide are mostly qualitative. The two methodologies are here illustrated with reference to Legewie et al. (2006) (ODEs) and Mai and Liu (2009) (LNs).

The bistable and irreversible features of apoptosis are well known. Bistability is thought to require a positive circuit, and so either a positive feedback or a double negative one. Beyond a certain threshold stimulus, such a circuit switches from OFF to ON state in an all-or-none fashion. The system displays hysteresis, i.e., different stimulus-response curves obtain depending on whether the system started in its ON or OFF state. Sometimes the ON state is maintained even upon removal of the stimulus, in an irreversible way. Legewie et al. (2006) offer a model of the intrinsic pathway of caspase activation (figure 3.1) that purports to show how the interaction between Casp3, Casp9 and XIAP (figure 3.1A) can result in a positive feedback which brings about bistability and helps generate irreversibility in the caspase activation.

The kinetics of the system comprise the reactions represented in figure 3.1B. Simulations were performed for the system of ODEs corresponding to such reactions. Results were derived by simulating response times of Casp3 activation upon variation in (total) active Apaf1. Upon weak stimulation Casp3 cleavage is slower, whereas upon strong stimulation it is faster, the response time being inversely related to the stimulus strength. Casp3 activity turns out, as expected, to be bistable and irreversible. The system has three steady states, two stable and one unstable, and shows hysteretic behaviour, having low active Casp3 for low active Apaf1 until a threshold point is reached where active Casp3 switches irreversibly to a high state.

Then, the authors investigate what is responsible for irreversibility in the presence of bistability. Other simulations showed the crucial role of XIAP (figure 3.2). Upon weak stimulation most Apaf1-associated active Casp9 is inhibited by XIAP, whereas above the threshold Apaf1 manages to initiate

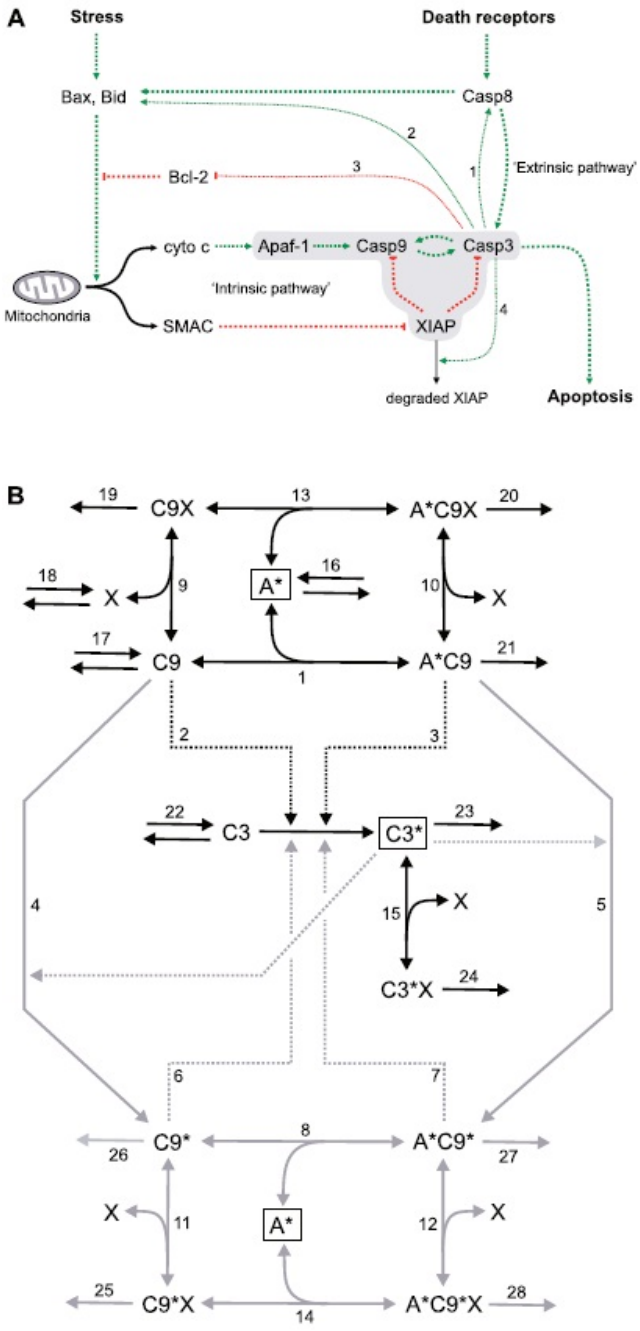

Figure 3.1: (A) Intrinsic and extrinsic pathways. Dotted lines indicate positive (green) or negative (red) regulation. The regulatory interactions considered in the model are highlighted in gray. (B) Kinetics in the model (X: XIAP; A*: activated Apaf1; C3: Casp3; C9: Casp9; C3*: activated C3*; C9*: activated C9*). Reproduced with permission from (Legewie et al., 2006, p. 1062).

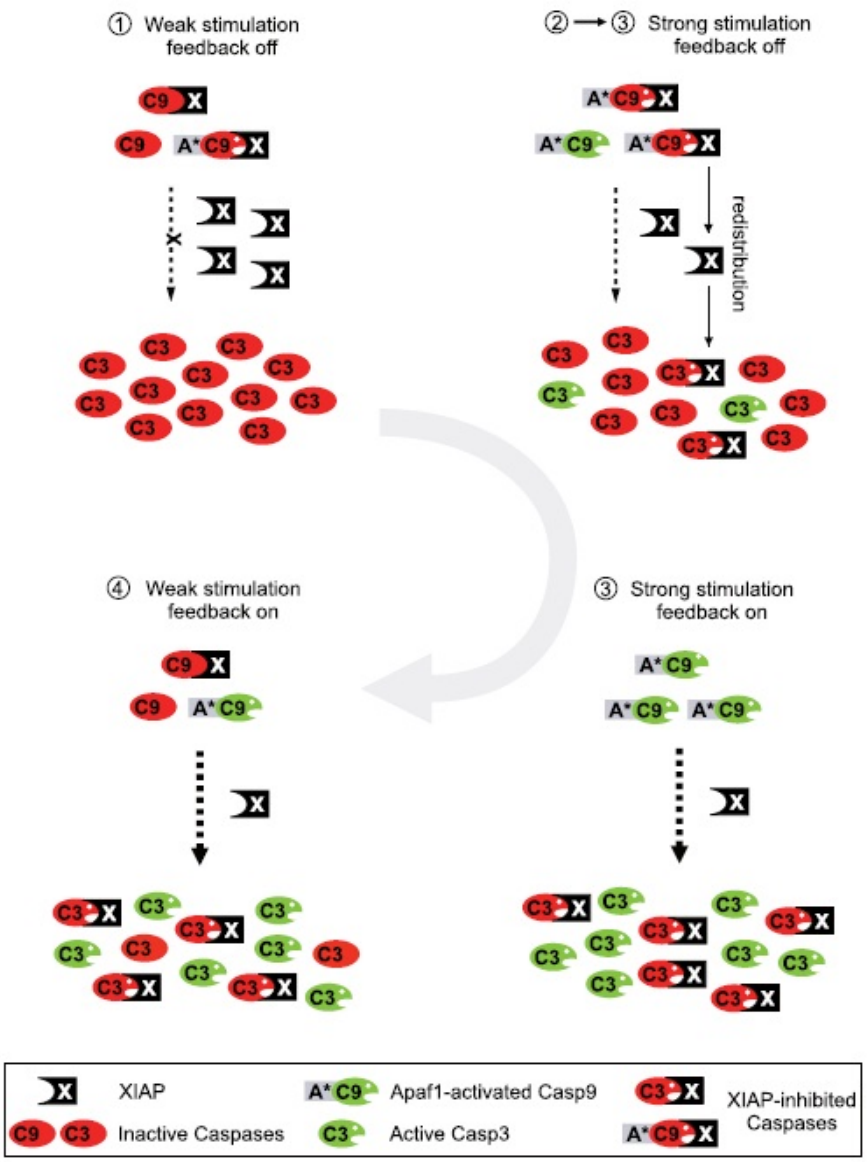

Figure 3.2: XIAP-Mediated Feedback. Top left: Casp9 is inhibited by XIAP, so that Casp3 is inactive. Top right: upon stronger stimulation some Casp9 escapes XIAP-mediated inhibition and activates Casp3, which then sequesters XIAP away from Casp9 (redistribution). Bottom right: redistribution results in strong activation of both Casp9 and Casp3. Bottom left: the system remains in an active state even if the stimulus is reduced. Reproduced with permission from (Legewie et al., 2006, p. 1066).

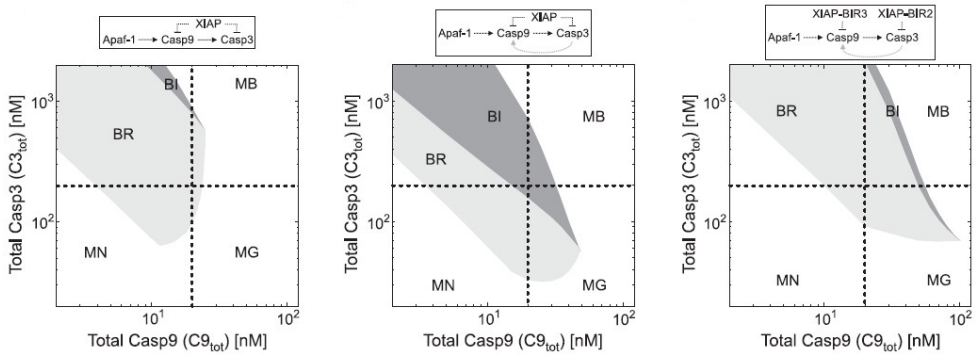

Figure 3.3: BI = bistable irreversible; BR = bistable reversible. Left: only XIAP feedback present. Centre: both XIAP and Casp3 feedbacks present. Right: XIAP is mutated, and binds to Casp3 and Casp9 in a noncompetitive way. Irreversibility depends on both (non-mutated) XIAP- and Casp3-induced feedbacks. Reproduced with permission from (Legewie et al., 2006, p. 1067).

Casp3 activation. Active Casp3 then further promotes its own activation by sequestering XIAP away from Apaf1-associated Casp9. Unless XIAP is mutated, so that it can non-competitively bind to both Casp3 and Casp9, this results in an implicit feedback: most of XIAP is bound to Casp3, so is unable to inhibit Casp9, which is then free to trigger executioner Casp3. Furthermore, Casp3 activity is maintained even if the stimulus is removed: once activated, Casp3 retains XIAP, thereby preventing full Casp9 deactivation. The authors conclude that the (non-mutant) XIAP-induced feedback is necessary, alongside the Casp3 positive feedback on Casp9, for the irreversibility of Casp3 activation (figure 3.3).

ODE models tend to focus on limited parts of the apoptosis network. A more complete picture involves not only both the intrinsic and extrinsic apoptotic pathways but also pro-survival pathways, e.g., the growth factor (GF) pathway. The epidermal growth factor receptors (EGFRs) are cell-surface receptors that, upon binding of their specific ligands, become active and trigger signalling transduction cascades that lead to DNA synthesis and cell proliferation. Mutations responsible for EGFR overexpression or overactivity can result in uncontrolled cell division, which is a predisposition for cancer. To understand the emergence of system properties in this more complex scenario, Mai and Liu (2009) built a 40-node LN (figure 3.4) and performed on it extensive statistical analyses. This resulted in the identification of key network components responsible for the stability of the surviving states and

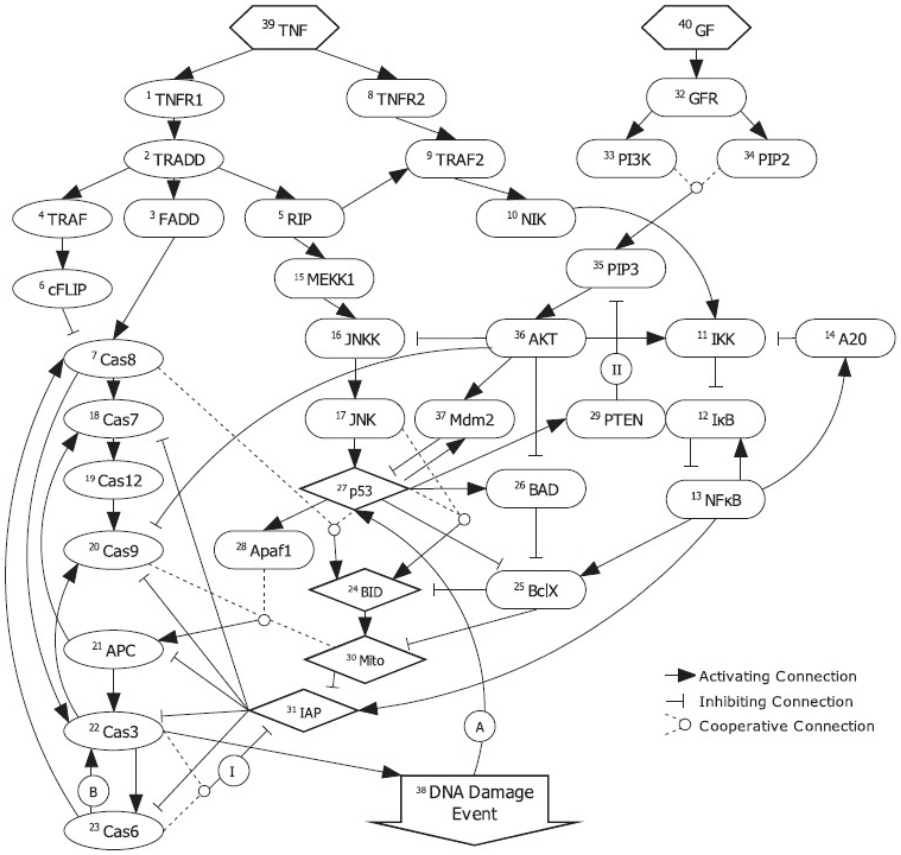

Figure 3.4: Boolean network in (Mai and Liu, 2009, p. 762), reproduced with permission. Each edge corresponds to inhibiting or activating connections. The cycles labelled A, B, I and II indicate connections removed in the knockout experiments. Reversing the states of nodes represented as cycles may result in survival-apoptosis transitions when GF is OFF. Nodes as diamond boxes may result in survival-apoptosis transitions when GF is ON.

the irreversibility of the apoptotic process.

The LN includes 37 internal nodes representing states of signalling molecules, 2 input nodes representing extracellular signals, and 1 output node corresponding to apoptosis ('DNA damage event'). Interconnections between nodes are based on literature and databases. The LN models a number of pathways and interconnections, such as extrinsic and intrinsic apoptotic pathways, the pro-survival effect of extrinsic TNF and GF signals, the regulatory connections at p53, the caspase machinery, and major links and feedbacks between pathways, e.g., the Bid-mediated crosstalk between signals from TNF and p53 and the negative feedback that the GF pathway receives from p53.

Each node can be either ON or OFF at each time step. All nodes apart from TNF and GF receive inputs from one or more nodes. The state of node i at $t + 1$, $S_i(t + 1)$, depends on its current state $S_i(t)$ and the magnitudes of the total activating inputs $A_i(t)$ and inhibiting inputs $H_i(t)$:

$$S_i(t+1) = \begin{cases} \text{OFF if } \mathrm{A}_i(t) < H_i(t) \\ \text{ON if } \mathrm{A}_i(t) > H_i(t) \\ \mathrm{S}_i(t) \text{ if } A_i(t) = H_i(t) \end{cases}$$

Simulations were run from 10,000 initial states randomly sampled. Each simulation run ends when either apoptosis or survival has been reached. This is judged according to whether the DNA damage event has been ON for 20 successive steps or has not taken place for 200 steps.

To find out key factors responsible for the *irreversibility* of apoptosis, the temporal evolutions were monitored under different combinations of input signals (both GF and TNF in their OFF state; only TNF ON; only GF ON; both GF and TNF ON) and interruptions of apoptotic signals (by setting and maintaining OFF selected nodes A, B, I and II, and combinations of them).

Results are obtained as regards the percentage of initial states leading to apoptosis over the total number of initial states (Apop%). The study of the varying dependence of Apop% on the external signal combinations confirms that TNF is a strong promoter of apoptosis. TNF effects, in fact, can be only partially offset by GF. The authors investigate whether withdrawing TNF or the mitochondrial signal after apoptosis has started has any effect on the irreversibility of the process. The results show that apoptosis is irreversible in both cases in the complete model. Data relative to models involving deletions of the four feedback connections, both separately and in combinations, by interruption of nodes A, B, I and II, show that feedbacks involving B and I (positive feedbacks containing Casp3) play more essential roles in promoting apoptosis than feedbacks involving A and II—although knockout of other combinations, too, can have an effect on—loss or degradation of—irreversibility.

The *stability* of the surviving states with respect to fluctuations in the states of the internal nodes is investigated, starting from the ending states, by systematically reversing the states of the internal nodes, either one-by-one or two at the time. The time evolution starting from each of the perturbed states is then followed until apoptosis or survival are, again, reached. There

are in total 40 final surviving states, associated with different combinations of input signals. Each surviving ending state was subjected to both single-node and dual-node perturbations. The authors study the number of perturbations leading to a survival-to-apoptosis transition for each surviving state. Their results suggest that although the GF signal does not significantly increase the overall survival ratio when there is no TNF signal, it can greatly increase the stability of the final surviving states in this case.

Such results are achieved at the price of large simplifications, e.g., components can only be in two states, the dynamics are governed by simple rules replacing complex molecular processes. Also, it must be noted that not all the initial states can be realised in real experiments, as the cell could not survive. Compared with ODEs, LNs are of limited power in approximating experimental results and making context-specific quantitative predictions. However, they allow for easier integration of experimental information and more systematic explorations of the state space. They also provide results that more readily offer themselves to a variety of statistical analyses.

3.2 Modelling economic complexity

3.2.1 Computational economics

Economics, as opposed to biology, has traditionally been more theory-driven than data-driven, meaning that general theoretical assumptions have tended to guide the generation of hypotheses. *Computational* economics is a branch of the emerging field of computational, or 'generative', social science. Its roots lie in a discomfort with mainstream, 'neoclassical' economic theory, both with its assumptions and its inability to account for certain empirical facts (Rickles, 2011; Hommes, 2006), discomfort which has grown stronger in the light of the recent financial crisis (Buchanan, 2009; Farmer and Foley, 2009).

Neoclassical economics is based on a set of implicit rules or assumptions (Weintraub, 1993): 1. People have rational preferences among outcomes that can be identified and associated with a value. 2. Individuals maximise expected utility and firms maximise expected profits. 3. People act independently and fully rationally (they use all available information and do not make systematic forecasting errors). It is assumed that non (fully) rational agents and firms will not survive evolutionary competition and will therefore be driven out of the market (see Hommes, 2006, pp. 1112-1113). These as-

sumptions concerning individual behaviour add up to the so-called 'rational expectation hypothesis' and give foundations to a general hypothesis about market behaviour, the 'efficient market hypothesis': prices always reflect all available information in actual markets, that is, they emerge aggregatively via the consensus amongst perfectly rational agents. Given that at each moment, in the light of available information, there is just one price that rational agents should agree upon, i.e., the rational price, the best estimate for the future price is the present price, since this price reflects all known information. For this reason, neoclassical economics focusses on explaining regularities by aggregating over the agents' rational behavior. At each time step, what is crucial for calculating an asset's price is the identification of the 'rational expectation equilibrium' (REE) reached by aggregation, that is, the solution to an agent maximisation problem. Any deviation from such a theoretical equilibrium is due to an exogenous intervention, i.e., a new piece of information, that changes the price *unpredictably*.

Neoclassical mathematical finance is based on the assumption that stock prices exhibit *geometric Brownian motion*. Let me explain.[31] Let us indicate time with t, $0 \leq t \leq \infty$, and the value of an asset at the present time t with $p(t)$. Let us define as (arithmetic) 'return' $R_\tau(t)$ the relative variation in an asset's price in the time interval t to $t + \tau$, i.e., the rate of price change:

$$R_\tau(t) = \frac{p(t+\tau) - p(t)}{p(t)} \tag{3.1}$$

The time interval τ can be taken as the difference between two successive trading days, or two periods when dividends are paid, etc. Neoclassical economics assumes that, for all non-negative values of t and τ, $R_\tau(t)$ is a random variable which is normally distributed and only depends on the present price. Given that most assets in finance are non-negative, the *log*-normal distribution is used to describe the probability density function of future prices. In general, a variable is lognormally distributed if its log is normally distributed. In our case, if returns are normally distributed, the future price is lognormally distributed (figure 3.5).

According to the standard model, if returns at t are independent of prices up to t and their log is a normal random variable, the prices $p(t)$ follow a geometric Brownian motion, or "random walk".[32] From this assumption

[31] See Ross (2003) for a more detailed exposition.

[32] The reason why the price sequence is called "geometric Brownian" is that each element in the sequence is obtained from the previous one by multiplying it by a certain factor, as in

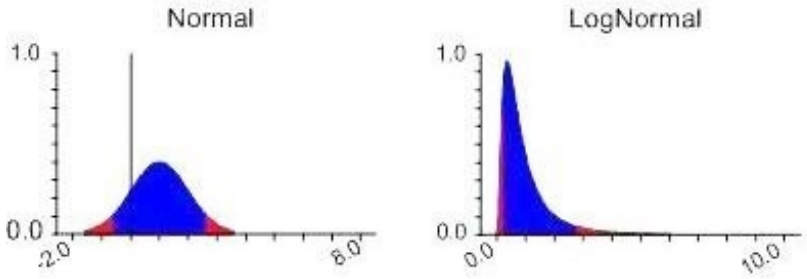

Figure 3.5: Normal (left) and Lognormal (right) distribution.

it follows that (i) the best estimation of an asset's future price is its current price; (ii) the distribution of price changes is Gaussian; (iii) purchases balance sales (cf. Rickles, 2011, §4).

However, the standard model goes only halfway in explaining financial complexity, as it cannot account for most of the following, well-known (often before the formulation of the standard model itself) empirical facts, also called 'stylised facts' (cf. Ross (2003, chap. 12), Rickles (2011, §6), LeBaron (2006, pp. 1191-1192), Samanidou et al. (2007, p. 411)).

Although unconditional distributions of returns at high frequencies (one day or so) are roughly Gaussian, which entails that the direction of stock returns is generally unpredictable, unconditional distributions of returns of assets at lower frequencies (one month or less) are *fat-tailed*, that is, have too many observations near the mean, too few in the mid range and too many in the tails to be normally distributed. Several, alternative distributions have been proposed to describe fat tails. Among them are power laws, which are commonly taken as a symptom of underlying complexity (see §1.3.4).[33]

Also, large (small) price changes tend to follow large (small) price changes, instead of being uniformly distributed (*volatility clustering*). Relatedly, asset returns at different times show a dependency (*volatility persistence*, or 'long memory'): whilst the autocorrelation of returns (i.e., the correlation between

geometric sequences (e.g., 1,2,4,8,16...), with the difference that, in a geometric *Brownian* motion, such a factor—which in our case is the return—varies randomly (in analogy with the sequence of random steps of the particles suspended in a gas, also called "Brownian"), whereas in traditional geometric sequences this factor is instead fixed.

[33]Power law distributions, characterised by the cumulative distribution function $\Pr(X > x) \sim ||x||^{-\alpha}$, can be used to describe not only stock price fluctuations but also trading volume and number of trades.

values of the returns at different times, as a function of their time difference) decays quickly to zero, providing support for the geometric Brownian motion hypothesis, the autocorrelation of *squared* returns decays more much slowly. Volatility persistence constitutes evidence of some predictability at longer horizons. Indeed, prediction methods based on such a predictability (e.g., the 'moving average technical analysis') are returning to fashion.

Finally, the observed *trading volume* is too high to be consistent with the efficient market hypothesis. A possible explanation is that volume is driven by differences in opinion between the agents, who perhaps are not so rational after all.

Computational economics offers the possibility to search the space of hypotheses on the mechanisms responsible for the stylised facts, and tries to assess their relative plausibility by suggesting some and discarding others. In general, computational economics tries to reproduce certain macrophenomena from the bottom up, starting from (more realistic) assumptions on the agents' microbehaviours. It mostly uses agent-based models (ABMs) (see §3.2.2), which is why it is often referred to as 'agent-based computational economics' (ACE).

ACE belongs to the wider, emerging, field of bottom-up computational, or 'generative', social science (Epstein, 1999, 2006). According to the generativist, it is not sufficient to establish that the system, once deposited in some macroconfiguration, will stay there. Guided by the motto 'If you didn't grow it, you didn't explain it', the generativist wants to account for how the configuration was attained by a *decentralised* system of *heterogeneous*, *autonomous*, agents. (Note, however, that the motto's converse doesn't hold: growing it isn't sufficient, only necessary, to explanation.) Statistics is then used to estimate the generative sufficiency of a given microspecification. The goal is to build models with empirically plausible rules that generate empirically adequate behaviours in order to *explain* the stylised facts.

ABMs are typically used to 'grow' the macrobehaviour. Several features differentiate ABMs from other modelling techniques (Epstein, 1999). The agents are *heterogeneous*, each with proper characteristics and preferences. They are *autonomous*, without central, top-down control over their behaviour. Their states and actions are modelled in an *explicit space*, or environment, whose topology is formally specified (e.g., a grid of cells, a network). They interact locally with their neighbours according to rules, whether merely reactive or also proactive, that produce behaviour as output of bounded infor-

mation and computing power.

ACE has several areas of application (Tesfatsion, 2002). When applied to the study of financial markets, ACE aims to establish microfoundations for the stylised facts, that is, to validate hypotheses regarding the *mechanisms* responsible for them.[34] Other obvious—although harder to obtain—*desiderata* are prediction and control over the real counterparts of the simulated facts (see, e.g., Buchanan, 2009).

Distinctions within the field can be drawn between a more physics-oriented approach (econophysics) and one more inspired to evolutionary biology (econobiology) (Rickles, 2011). Accordingly, the economic system can be envisaged as a self-organising system—along the lines of Nicolis and Prigogine (1989); Bak (1997); Sornette (2002)—or as a complex adaptive system—along the lines of Holland (1995) and Casti (1997). In §3.2.3, two ABMs of asset pricing will be presented, one inspired by the econophysics approach, another by the econobiology approach.

3.2.2 Asset pricing

The stylised facts (fat-tailed distributions of returns, high volatility, large trading volumes, etc.) demand that we re-interpret the economic agent not as *perfectly* rational, that is, capable of taking optimal decisions in the light of full knowledge of facts, but as *boundedly* rational (Simon, 2000). This means, in short, that the agent starts with a set of forecasting hypotheses, none of which necessarily correct, and then tests and changes them *inductively*. In the market, agents' decisions depend on bounded rationality. Some traders continuously decide whether to buy or sell by trying to identify price trends and patterns and guessing what other traders will do; their decisions then influence the market, which in turn influences their future decisions. Given the self-referential nature of this process, the agent is never in the position to deduce what the best decision has to be. Macro-equilibrium, when attained, emerges 'ecologically' rather than as a result of deductive reasoning.

Soros (1987) labels this phenomenon "self-reflexivity". In general, reflexivity refers to allegedly circular relationships between cause and effect. The reflexive relation links the thought of the actors involved to the situation they are part of, either developing toward the equilibrium or generating changes

[34] For reviews of ABMs of financial markets, see Hommes (2006); LeBaron (2006); Samanidou et al. (2007).

and inverting trends. Soros applies this concept to the mutual relations between the course of the market and its participants' expectations.

The resulting theory is in opposition with equilibrium theory. The latter, in fact, stipulates that markets move towards equilibrium and that fluctuations are merely random noise that tend to be corrected quickly. Let us define the intrinsic, or *fundamental*, value of an asset (FV) as the discounted sum of future earnings. This is calculated by summing the future income generated by the asset (interests), and discounting it to the present value:

$$\mathrm{DPV} = \sum_{t=0}^{n} \frac{\mathrm{NV}_t}{(1+i)^t} \tag{3.2}$$

where DPV stands for the discounted present value of the future cash flow, or FV adjusted for the delay in receipt; NV is the nominal value of a cash flow amount in a future period; i is the interest rate; and t denotes the time periods where cash flow occurs. In equilibrium theory, long-run prices reflect the underlying fundamentals, the allegedly 'real' values of the assets, which are unaffected by current prices.

The theory of reflexivity, instead, states that prices *do* influence fundamentals and that the influenced fundamentals then change expectations, thereby influencing prices in a self-reinforcing pattern. So, for instance, if traders believe that prices will fall, they will sell, driving prices down, whereas if they believe prices will rise, they will buy, driving prices up. Reflexivity is thus related to feedback mechanisms. Since the pattern is self-reinforcing, markets tend towards disequilibrium. At a given point, positive (negative) expectations overcome negative (positive) ones and become self-reinforcing in the upward (downward) direction, which explains the familiar pattern of boom-bust cycles.

ABMs can help bridge the gap between individual behaviour and its collective outcomes by, among other things, providing a representation of the traders' bounded rationality in an attempt to explain the stylised facts. Simon's idea of of bounded rationality has been adopted, at least implicitly, by Arthur et al. (1997), whose model of asset pricing is described in §3.2.3, and in general by those who envision the market as an evolutionary-adaptive system, whose participants face self-referential decision problems (cf. Markose et al., 2007). Also, although Simon is rarely cited in the econophysics literature, the intuition that the target of the research should be equilibria *formation*, starting from a study of the dynamics of agents who lack rational expecta-

tions and optimisation principles, is widespread. Accordingly, the economy can be re-interpreted as an out-of-equilibrium system, where self organisation can emerge spontaneously, without external interventions, and lead to one or the other among *several* possible equilibria, sometimes switching from one to the other when small amounts of noise are introduced (Arthur, 2006).

There is no generally accepted model for the formation of the expectations of the economic agents. Different models assume different behavioural rules to reproduce the stylised facts. One way is to group agents into different categories, each with a different attitude towards investing, and explain fluctuations in the market by reference to agents switching from one category to another with certain probabilities in response to the situation—as in the model in (Lux and Marchesi, 1999, 2000). An alternative way is try to model the agents's learning process *directly*, to mimic the way in which they inductively adapt their trading strategies—as in the model in (Arthur et al., 1997) and (LeBaron et al., 1999). Usually, 'genetic algorithms' (GAs), first introduced by Holland (1995), are used for this latter task. GAs are employed to solve *optimisation* problems, whose task is to find not solutions that maximise some utility function common to all agents, but 'satisficing' solutions, i.e. solutions that are 'good enough' given the agents' limited computational ability and access to information (Simon, 1996, chap. 2).

The first step of the GA modelling procedure is to map the behavioural rules into a genetic structure, e.g., by coding real-valued parameters as strings of 0's and 1's. Attached to each gene in this population is a fitness value, representing how well the rule, or solution, has performed in solving the optimisation problem. The computer then simulates evolution by searching the space of the possible genotypes for those of high fitness: it does so by creating new genotypes from old and evaluating the relative fitness of each genotype in the population (see Casti, 1997, pp. 158-161). This involves *mutation* (random flip of bits in the strings of 0's and 1's), *crossover* (interchange of subsequences of two genotypes to create two offspring) and *selection* (fitter genes are more likely to reproduce than less fit ones, so that a new population with higher-fitness solutions tends to replace the old one). A run of the GA can consist in hundreds or thousands of generations, after which one or more very fit genes are selected, which are then taken to constitute a 'solution' to the original optimisation problem.

In the case of financial settings, genotypes represent agents, or 'theories' of the market, whereas single genes stand for trading rules, or strategies.

Theories evolve and are evaluated on the basis of their performance. Rules are formed depending on time series information.[35] Rules are then assessed according to how well they, e.g., minimise some forecast-based error measure. The best performing rules are used to formulate the agents' forecasts, which are in turn converted into asset demands using preferences. Finally, new rules are generated by means of a GA. By using this procedure, it is possible to evaluate and compare agents based on their forecasting performance.

3.2.3 Modelling artificial stock markets

The first model I describe originates from the study of many-particle systems in physics and is presented in (Lux and Marchesi, 1999, 2000). The model describes an asset market with a fixed number of traders which, contrary to the rational expectation hypothesis, are divided into two main groups: *fundamentalists*, who sell (buy) when the price is above (below) the fundamental value, and *chartists* (or 'technical traders', or 'noise traders'), who buy or sell depending on the other traders' behaviour and the prevailing price trend. Chartists, in turn, are subdivided into optimistic and pessimistic. In short, the system works as follows: traders can switch between different groups; the number of individuals in these groups determines the excess demand, i.e., the difference between demand and supply; imbalances between excess and demand result in changes in actual price, which in turn affect the agents' trading strategy. The dynamics of the model are determined by four components: the chartists' switch from pessimistic to optimistic behaviour, and vice versa; the traders' switch from fundamentalist to chartist behaviour, and vice versa; the actual price changes as a result of the endogenous responses to demand-supply imbalances; changes in fundamental value, governed by a random process, so as to assure that the resulting stylised facts do not depend on exogenous factors.

The probability of switches among chartists is governed by the development of two noise factors influencing the switch, that is, an *opinion index*, representing the average opinion among chartist traders, and the *price trend*, i.e. the variation of the actual price in time. Two parameters in the probability function regulate the sensitivity of the traders to, respectively, opinion and price trend. The probability of switches between chartists and fundamentalists depends on the difference between the momentary profits earned

[35] In §3.2.3 I describe a method known as 'classifier system', used by Arthur et al. (1997) to convert time series information into trading rules.

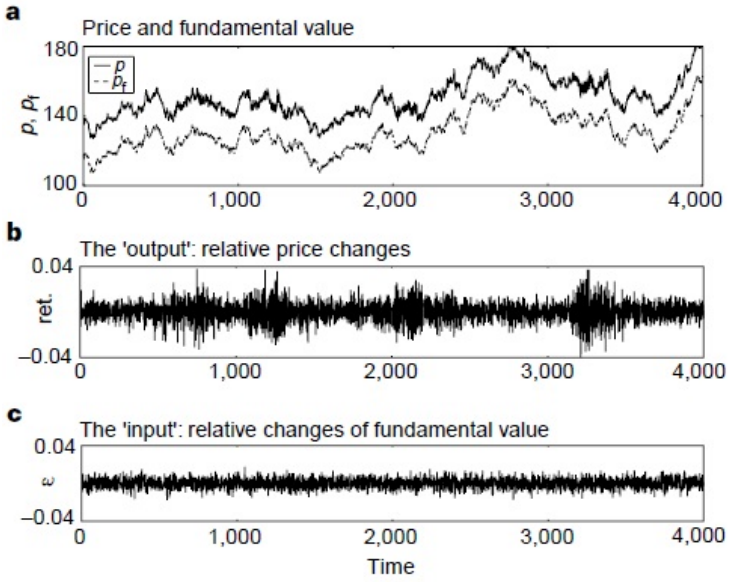

Figure 3.6: Time series of (a) market price and fundamental value, (b) logarithmic returns (ret), i.e. log changes of the market price: $ret_t = ln(p_t) - ln(p_{t-1})$, and (c) log changes of the fundamental value: $\epsilon_t = ln(p_{\mathrm{f},t}) - ln(p_{\mathrm{f},t-1})$. Reproduced with permission from (Lux and Marchesi, 1999, p. 498).

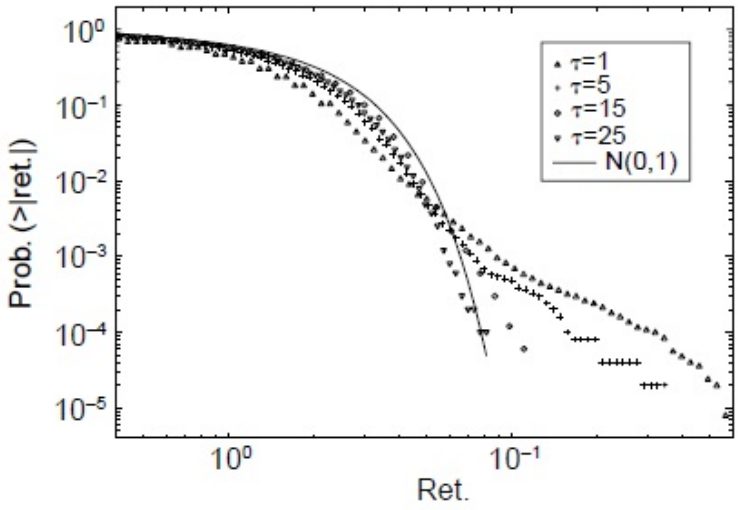

Figure 3.7: Log-log plot of the complement of the cumulative distribution of returns at different levels of time aggregation: $ret(\tau) = ln(p_\tau) - ln(p_{t-\tau})$. The distribution's decay is close to the exponential decay of the Normal at low frequencies (viz. with larger time interval τ), whereas it approximates power-law scaling at high frequencies. Reproduced with permission from (Lux and Marchesi, 1999, p. 499).

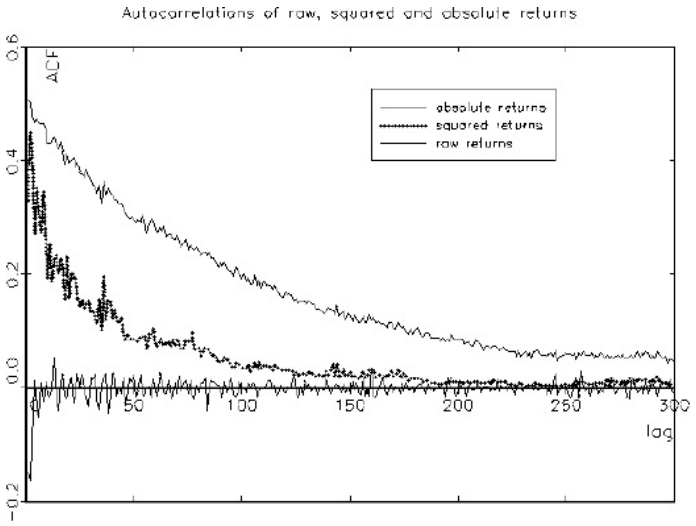

Figure 3.8: Autocorrelations of raw (bottom), squared (middle) and absolute (top) returns. Reproduced with permission from (Lux and Marchesi, 2000, p. 695).

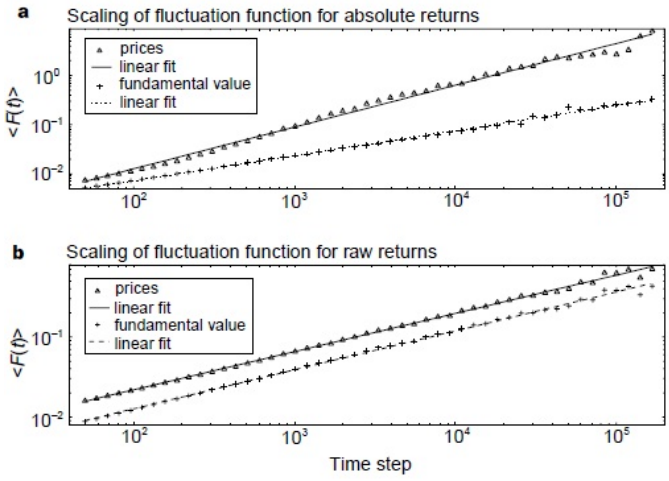

Figure 3.9: Comparison between the scaling of the fluctuation function of changes in fundamental value and the scaling of the fluctuation function of, respectively, raw returns (bottom) and absolute returns (top). Reproduced with permission from (Lux and Marchesi, 1999, p. 499).

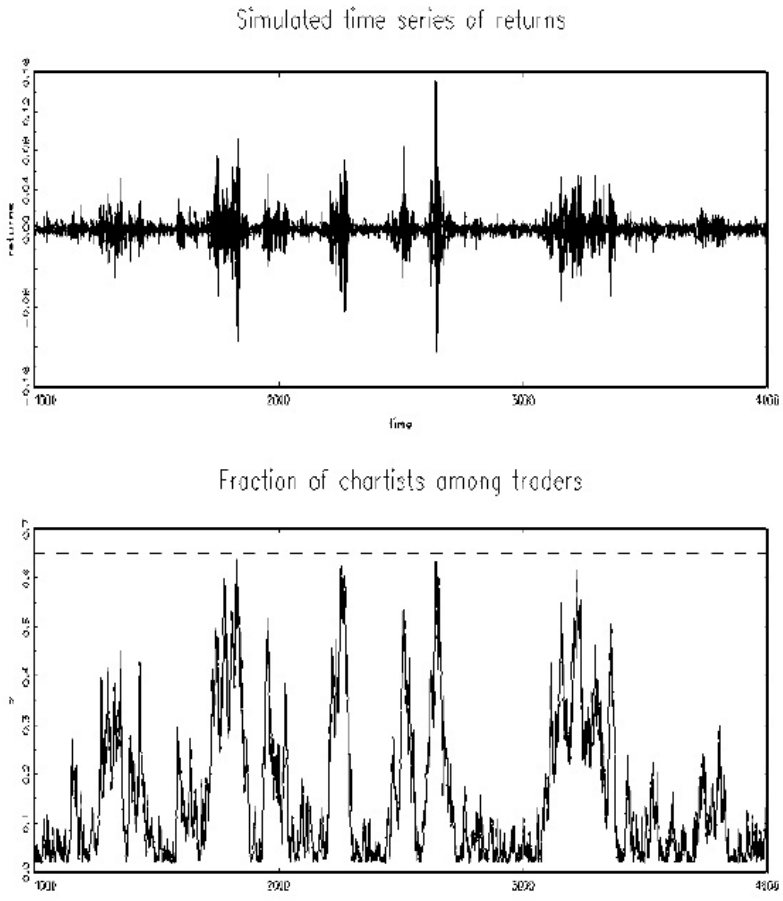

Figure 3.10: Dependence of time series of returns (top) on proportion of chartists in the market (bottom). When the fraction of chartists exceeds a critical value, the system tends to become unstable. Reproduced with permission from (Lux and Marchesi, 2000, p. 689).

by individuals in both groups. Chartists regard as profit the *realised* excess profits, that is, short-term capital gains due to the price change. Fundamentalists, instead, regard as profit the *expected* excess profits, that is, the difference between price and fundamental value, which they take as a source of arbitrage opportunity, to be realised when the future price reverts to the fundamental value. A parameter in the probability function specifies the sensitivity of the traders to the profits. Price changes are determined by a *market maker*, who adjusts the price so that the variation per time is proportional to the aggregate excess demand of chartists and fundamentalists.

Of the three types of steady states of the system, the one of interest obtains when the price is at its fundamental value, there is a balanced proportion of optimists and pessimists, and an arbitrary proportion of chartists. This steady state is repelling (i.e., the system is unstable) when either (i) the parameters measuring sensitivity to opinion, price changes and profits are larger than some critical value or (ii), if such parameters are below the critical value, when the corresponding proportion of chartists exceeds a critical value. Simulations are performed for parameter values for which the system is in the repelling steady state, and show how an otherwise stable equilibrium can be subject to transient phases of destabilisation, with fluctuations around the equilibrium that suddenly emerge and quickly die out. This 'punctuated equilibrium' generates time series with clusters of excessive volatility interspersed among long tranquil periods.

The time series of the market price stays close to the time series of the fundamental value, in agreement with the hypothesis that prices follow a random walk (figure 3.6a). Still, statistical analyses show the presence of 'abnormal' features: *contra* the standard model, the normally distributed log changes in fundamental value (and the absence of exogenous shocks) (figure 3.6c) do not result in similarly normally distributed returns, the time series of returns exhibiting a higher-than-normal frequency of extreme events and volatility clustering (figure 3.6b).

A study of the complement of the cumulative *unconditional* distribution of returns (figure 3.7) shows how the probability of large fluctuations doesn't depend only on current price but also on the time interval considered. At high frequencies (with small τ) the distribution approximates a power-law. At low frequencies (with large τ), instead, the distribution decays quickly, approximating the exponential decay typical of the Normal. This agrees with the observation of financial prices obtained, respectively, at high and low

frequencies: large price fluctuations at high (daily) frequencies are scarce, and the tails of the corresponding distributions behave normally; large fluctuations at low (weekly, monthly) frequencies, instead, are more numerous than Normal, with the tails being better described by a Pareto distribution, $\Pr(X > x) = ax^{-\alpha}$—which is a power-law distribution.

The study of the *conditional* distribution of returns shows volatility clustering: against the view that deviations from equilibrium are unpredictable, periods of quiescence and turbulence tend to cluster together. Use of elementary statistical techniques shows that whereas raw returns have low autocorrelation and fluctuate around zero, which is indicative of short memory, squared and absolute returns show much higher autocorrelation, which indicates long memory (figure 3.8). More sophisticated techniques, viz. 'detrended fluctuation analysis' (see Kuhlmann, 2011, §3.1), are employed to study the scaling properties of the average fluctuations $F(t)$ of fundamental values, raw and absolute returns as a function of the time interval t, and to calculate the exponents of the corresponding power laws (figure 3.9). The analysis reveals that the slope of the power law of changes in fundamental price is similar to the slope of the power law of raw returns, but smaller than that of the power law of absolute returns, which is a sign of strong persistence in volatility.

The authors conclude that, since these scaling properties are absent in the behaviour of the exogenous force, viz. the changes in fundamental value, they are endogenously generated by the interaction (switching) of heterogeous economic agents. In particular, the authors notice that for a wide range of parameter values the volatility bursts robustly depend on whether or not the proportion of chartists in the market exceeds a critical value (figure 3.10). Hence, they conclude that the volatility bursts are *explained* in terms of the switches between groups driven by chartist behaviour (see Lux and Marchesi (1999, p. 500) and Lux and Marchesi (2000, p. 679)). The resulting punctuated equilibrium is considered analogous to the on-off intermittency found in many natural systems, where an attracting state may become temporarily unstable due to a local bifurcation until the system is driven back to stability by some endogenous mechanism. The bifurcation obtains when some key variable surpasses some stability threshold—in the present case, this is the time-varying fraction of chartists. Interestingly, since intermittency manifests itself through features proper of chaotic phenomena, such as the scaling behaviour of the time intervals of tranquil periods inbetween severe fluctuations (cf. Smith, 1998, pp. 110-111), finding such scaling properties in real time

series provides some evidence that the stock market is chaotic as assumed by the model.

I pass now to describe another model of asset pricing, the so-called 'Santa Fe artificial stock market' (Arthur et al., 1997; LeBaron et al., 1999). In the market, there is a fixed number of traders who determine endogenously the price of a stock by aggregation of their demand for the stock, which is directly proportional to their expectation of future wealth and inversely proportional to the variance of the asset's price and dividend. The homogeneous REE is used as a benchmark to evaluate the agents' individual strategies and calculate the clearing price. The model diverges from classical equilibrium economics in that the expectation formation is heterogeneous rather than homogeneous.

Each agent forms his expectations of the next period's price and dividend *individually* and *inductively*, by observing the state of the market (which includes the historical dividend and price sequence) and continually revising his 'theory' of the market in order to obtain better and better predictions. Each theory is made of 100 strategies, or rules. The learning process is modelled by means of a *classifier system* that codes rules into sets of predictors, each consisting in a condition part (a bit string of market descriptors) and a forecast part (a parameter vector)—which is why this classifier system is also referred to as *condition/forecast* classifier (cf. LeBaron, 2002). The condition part of a predictor j contains a 12-bit string of 1's or 0's, which can be interpreted as the current price, respectively, fulfilling a market condition and not fulfilling the condition. Each bit represents either technical or fundamental information. The forecast part contains parameters a_j and b_j of a linear forecasting model, namely the expectation function associated with the predictor j, as a linear combination of price and dividend, $E_j(p_{t+1} + d_{t+1}) = a_j(p_t + d_t) + b_j$. The forecast part also contains an estimate of j's current variance, σ_j^2.

At the start of the time period the current dividend is posted and observed by all agents. Each agent checks which of his predictors are 'active', i.e., match the current state of the market. He then forecasts future price and dividend by combining statistically the linear forecast of the most accurate of his active predictors, and given this expectation and its variance calculates the asset demand and makes the appropriate bid or offer. A price is then determined for clearing the market. After market clearing, the new price and dividend are revealed and the accuracies of the active predictors are updated. At regular intervals, but asynchronously, agents engage in a learning process for updating their set of rules. When learning takes place, the 20 worst

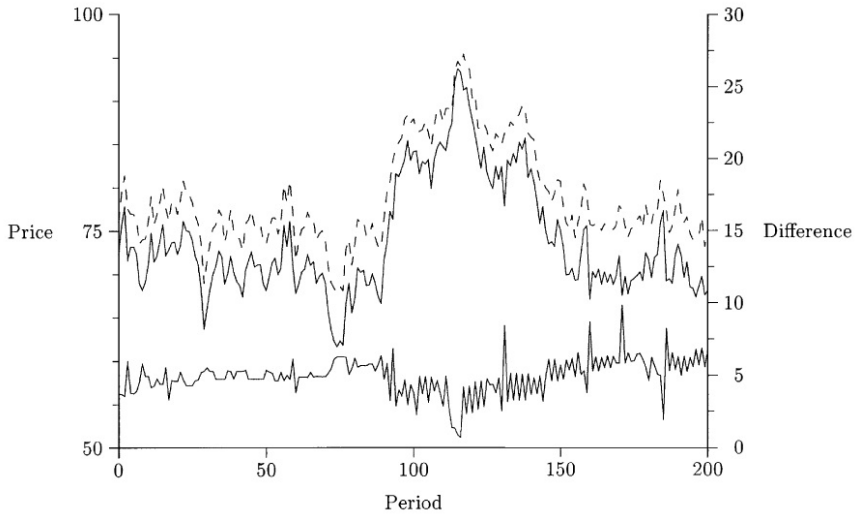

Figure 3.11: Time series of actual price (solid line), theoretical price (dotted line), and difference between the two (bottom line). Reproduced with permission from (LeBaron et al., 1999, p. 1500).

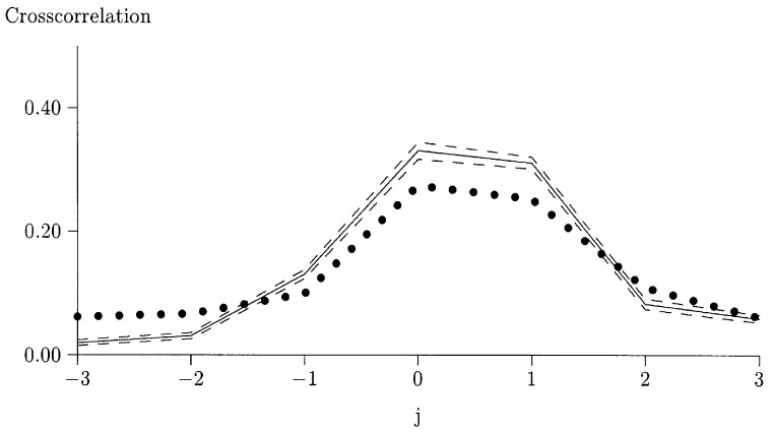

Figure 3.12: Correlation of squared returns at $(t+j)$ with trading volume at t. Solid line: mean fast learning; dotted line: comparison series (IBM daily returns, period 1962-1994). Reproduced with permission from (LeBaron et al., 1999, p. 1505).

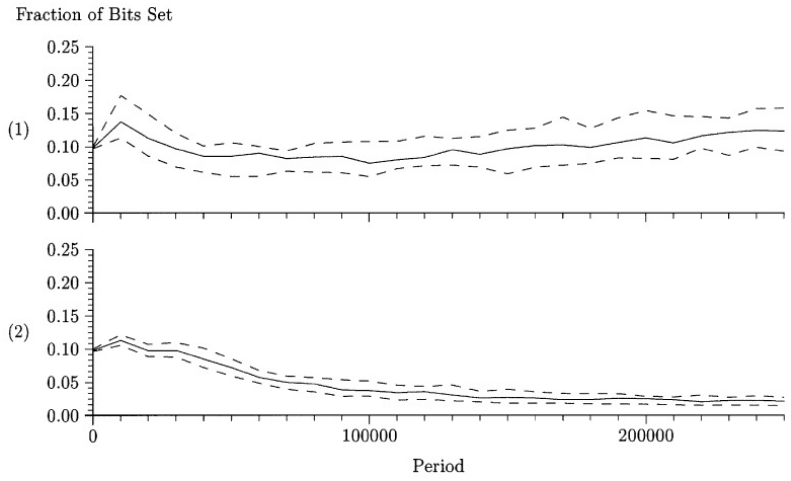

Figure 3.13: Technical trading bits under fast learning (1) and slow learning (2). Reproduced with permission from (LeBaron et al., 1999, p. 1506).

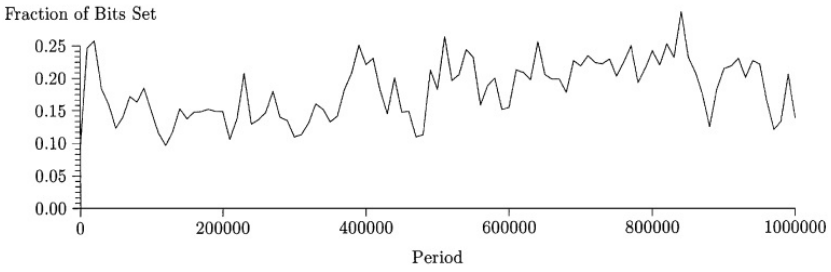

Figure 3.14: Long-run oscillations of technical trading bits. Reproduced with permission from (LeBaron et al., 1999, p. 1507).

performing rules are eliminated and replaced by new rules formed by means of a GA with uniform crossover and mutation.

Simulations were run under a variety of conditions. Most importantly, different parameter values ensuing in different rates of the learning process were used. Statistical analyses were then performed to study the properties of the resulting time series (e.g., predictability, trading volume, volatility) and the features of the learning process that determine them (condition bits used, convergence of forecast parameters). The general conclusion was that with slow learning the price series are indistinguishable from what should be produced in the case of homogeneous REE, whereas with fast learning several stylised facts are endogenously produced.

First, the actual price series, calculated by letting every agent adjust his demand according to his own forecasting rules, tracks very closely the theoretical price series, where market prices are calculated assuming homogeneous agents. Still, the time series of differences between actual and theoretical price shows the presence of both tranquil periods and wild fluctuations (figure 3.11). This result is qualitatively very similar to that obtained by Lux and Marchesi: the price time series contains evidence of underlying chaos.

Secondly, there is evidence of predictability at large horizons: forecasting models taking into account information that should be of no use in the REE (e.g., 500 period moving averages), are more successful than simple linear forecasting models based on the last period's price and dividend (LeBaron et al., 1999, p. 1503).

Thirdly, autocorrelation of volatility and trading volumes is lower with slow learning and higher with fast learning, in this latter case approximating more closely values from observed time series. Also, volatility and trading volumes are strongly correlated in the case of fast learning, again approximating observed values (figure 3.12).

A study of the evolution in the use of the condition bits shows that with slow learning the traders learn that technical bits are of no use and as the time progresses tend to eliminate them from their trading strategy. With fast learning, instead, the average use of technical bits does not decay (figure 3.13). Also, there is some evidence that in the long run, although the use of technical bits eventually stops increasing, it keeps oscillating (figure 3.14). This marks a potential difference between the Santa Fe market and Lux and Marchesi's market: whereas the latter conforms better to Sornette's model of on-off intermittency, viz. the system's dynamics are responsible for both

chaotic and non-chaotic periods, the former fits better Bak's model of self-organised criticality, viz. complex regimes are the stable, 'ordered' result of a system following the path of one, chaotic attractor.

The authors take these results to show that complex regimes may arise even under neoclassical conditions and in the absence of exogenous shocks. They explain this fact by reference to the heterogeneous and inductive nature of the agents' expectation formation process, ultimately responsible for fast learning. They notice that when technical trading strategies emerge they tend to become mutually reinforcing: trend-following strategies are randomly generated in the population; if random perturbations in the dividend sequence validate them, they then take their place in the population and change the market; the changes manifest in the form of increased volatility and volume and trigger further changes in strategies; this ensues in a complex, mutually reinforcing behaviour.

3.3 Simulating causal facts

Systems biology and computational economics are sciences that produce and use their results based on computation and/or simulation of more or less fictional scenarios whose physical realiser is a computer hardware rather than the system itself. Since the methodologies of systems biology and computational economics constitute a somehow novel way to test and use scientific hypotheses, some preliminary discussion on the differences between simulating causal scenarios and experimenting on them is needed before we can proceed to discuss to what extent accounts of causality capture the meaning of causal claims as derived from the models.

Simulation has both disadvantages and advantages with respect to traditional experiments. The major *disadvantage* commonly associated with simulation is that the material which is observed and manipulated is a model rather than the system itself.

Some (see, e.g., Guala, 2012, p. 611) argue that because of this, although a simulation may 'surprise' us, leading to results that we hadn't realised were implicit in the premisses of the model, it cannot 'teach' us anything new. The reason is that, contrary to the design of an experiment, which is partly 'opaque' due to a lack of knowledge of the inner workings of the experimental system, the design of the model is transparent to the modeller. But this argument is debatable. Humphreys, for instance, identifies in the

'epistemic opacity' one of the distinctive features of simulations—no human user can actually understand the process leading from the abstract model to the output (see Humphreys, 2009, pp. 618-619).

Others, instead, associate the materiality problem with the *reliability* of the inferences, the idea being that if the system studied is of the 'same stuff' as the target system, it is also conducive to more reliable inferences. However, as noted by Frigg and Reiss (2009), this is not necessarily the case. There are many theoretical models applied to experimental systems that are more reliable than experimental systems applied to non-experimental systems. In particular, materiality makes no distinctive feature—hence no disadvantage—for simulations vis à vis traditional experiments.

Be that as it may, simulation also offers undeniable *advantages*, insofar as they facilitate inferences in cases where experimental results aren't available. Epstein (2008) lists many advantages of simulation, besides prediction. Interestingly, many of them may be read as having a *causal* significance: simulations "explain" (C is a possible cause of E); they "suggest dynamical analogies" (C may cause E by a process analogous to that by which X causes Y); they "bound (bracket) outcomes to plausible ranges" (C may cause E, but only $E' < E < E''$); they "challenge the robustness of prevailing theory through perturbations" (C may cause E only across a given range of parameter values); they "expose prevailing wisdom as incompatible with available data" (C could not possibly cause E); they "reveal the apparently simple (complex) to be complex (simple)" (E may have multiple causes/E may have a unique cause); etc.

Sometimes scientists stress even more explicitly the causal significance of simulations. They take simulations to provide causal explanations (Kuipers, 1987). They use them to answer counterfactual questions on qualitative long-run policy effects and to identify mechanisms relevant to policy decisions (Dawid and Neugart, 2011). They use them to draw causal conclusions in cases of non-linearities, feedbacks, heterogeneity and adaptivity of the agents, various levels of organisation, etc. (Galea et al., 2010). The question, then, is not *whether* simulations may in principle deliver causal conclusion but *in what conditions* one becomes entitled to such conclusions.

Simulations' conclusions are, like those derived by other methods, only as safe as the premisses from which they are derived.[36] Assessing the goodness

[36]Those who believe that simulations are deductions (see, e.g., Epstein, 1999, 2006) will maintain that the conclusion is only as safe as the *weakest* of the premisses, considered in

of the derivations amounts to assessing the 'validity' of the inferences. This involves evaluating the warrant for the premisses, the need to include missing premisses, the relation between premisses and conclusion, etc. In particular, the advantages and disadvantages of simulations must be assessed with respect to the two distinct issues of internal and external validity. In scientific parlance, 'internal validity' refers to the correctness of the methodology by which the hypothesis is tested (or supported, or confirmed); 'external validity', instead, refers to the exportability of the results obtained in the test situation to some target situation. Before turning to issues of internal and external validity of simulations, some preliminary introduction to the way simulations are built and used is needed.

The simulation methodology consists of the following steps (see, e.g., Gilbert and Troitzsch (2005, chap. 2), Macal and North (2005)): 1. definition of the target; 2. gathering of observation of the target for getting parameters and initial conditions for the model; 3. making assumptions; 4. designing the model in the form of a computer programme; 5. running the simulation; 6. verification (i.e., debugging of the programme); 7. validation (ensuring simulated behaviour matches data collected from the target); 8. sensitivity analysis (via randomisation over parameter values and exogenous factors, changes in the order in which the actions are performed, analysis of the statistical features of the outcomes of the simulations, etc.).

Steps 6 through 8 raise issues of validity of the simulation, i.e. whether it succeeds in representing or reproducing some actual or counterfactual scenario. Verification involves internal validity issues (problems with, e.g., the approximation of numerical solution to actual solutions, truncation errors, the pseudo-randomisation involved in the sampling process). These are technical problems, which add to traditional internal validity problems with experiments on real systems. Validation, instead, involves both internal and external validity issues. Since the test is, strictly speaking, performed on a computer, the validity of its results as applied to some 'target' always involves an extrapolation, although one may distinguish between an 'internal extrapolation', where one's aim is to match the behaviour of simulation and the test system that was initially observed in steps 1 through 3, and an 'external extrapolation', where one's aim is to apply the model to non-actual scenar-

isolation (cf. Cartwright, 2007b, chap. 3). Instead, those who believe that the support the premisses lend to the conclusion depends partly on their number and mutual relation, as in inductive or abductive arguments, will judge the conclusion only as safe as the *conjunction* of the premisses (see, e.g., Winsberg, 1999, 2009).

ios, viz. the same kind of system under different circumstances, e.g., after an intervention. There may be several difficulties here: the stochasticity of both target and model; the sensitivity of the behaviour to chosen parameter values and initial conditions; the selection of data about the target to validate the model; etc. Finally, sensitivity analysis aims to overcome both internal validity problems (missing data on test system/population) and external validity problems.

3.3.1 Internal validity

Since the model is, strictly speaking, the object of experiment, one has here the additional problem of verifying that there is no mismatch between model and test system. As pointed out by Rickles (2009), there is a disanalogy between experimenting on a system and experimenting on a simulation: in the former case, we are speculating on whether the intervention may be used to validate a causal hypotheses; in the latter case, instead, we are using the simulation to model the dynamics of a complex system, "in a bid to understand what kind of process might have generated some data" (Rickles, 2009, p. 89). Although Rickles admits that this does not rule out other potential uses, he claims that "to get a simulation going we must have a causal model in hand. However, this is precisely what we were aiming to establish, so we will have reasoned in a circle adopting a simulation-based response" (*ibid.*). But this criticism seems too harsh. We normally need causal assumptions to establish causal conclusions. Provided we are not assuming the very same relation we are trying to establish, the 'circle' won't be vicious.

True, since—by design—the conclusion depends on an explicit set of premisses, one can't use the model as a a 'black box' that will, by some partially unknown mechanism, produce the same results in test system and target system in virtue of their similarity. Rather, one must ensure, through validation, that the internal design (principles, rules, laws, etc.) is correct. This is generally hard. If the design is too realistic, one needs a vast amount and variety of data. If it is abstract, it will be difficult to interpret and measure in the target the quantities (variables, parameters, etc.) present in the model (cf. Cartwright, 2007b, pp. 38-40).

Yet, since the model was explicitly designed, it is transparent to the modeller. With more realistic models, it is easy to interpret the conclusions in relation to the target, and to judge whether the model is 'faithful'. With more

abstract models, instead, one can draw more general conclusions, usually at the price of a more difficult choice of the data that can validate the model. However, the transparency of the model also makes it easier to test for the robustness of the results by means of sensitivity analyses. So, one can still say that something general is true of a large class of systems, to the extent that the results do not depend on precise parameter values, initial conditions, unrealistic assumptions, etc.

A further problem that Rickles (2009) sees in the use of simulations to establish causal conclusions is that, contrary to complex systems, which are always open to and entangled with the environment, simulations are in a clear sense closed, i.e. screened off from exogenous causes not specified by the model, but which have in reality the potential to amplify or modify causal effects. But openness and contextuality make problems for *any* attempt to draw conclusions with the aid of experiments, and do not affect simulations only. Also, openness and contextuality make it particularly difficult to *export* results rather than *establish* them. This is more a matter of external validity rather than internal validity. In this regard, *any* model may represent well some test situation but then fail to be exportable to some target situation.

3.3.2 External validity

All methods face the problem of external validity, "since we seldom establish results in the very population and in the very situation in which we want to apply them" (Cartwright, 2007b, p. 36). In particular, any experiment on some test population/system is similar to the target population/system only in certain respects and degrees. When external validity is concerned, material similarity is not necessarily more important than other kinds of similarity. For instance, analogous models, such as the Phillips curve fluid mechanics model of the economy that one observes in the Science Museum of London, are not materially similar to their targets, and yet can adequately reproduce some of their features. Simulations count as experiments in that, when it comes to export their results, problems of external validity arise as with any other method (physical models, analogous models, equation-based models, etc.) (cf. Frigg and Reiss, 2009, p. 597).[37]

To warrant the exportability of results from the experimental system to the target system, one should in principle make sure that no confounders and

[37] Also Guala concurs on this (see Guala, 2012, pp. 610-611).

no difference in causally relevant respects are present (Guala, 2012, §4.3). In traditional experiments, one relies on good subject sample and design. If there are problems with the former, one usually randomises over the test population. If there are problems with the latter, one may randomise over diverse situations, where arguably different tendencies and mechanisms are at work. Alternatively, one may look for the presence of marks of crucial stages, or 'marks', of the mechanism in both test and target (Steel, 2007, chap. 8) and/or ensure that test results depend on tendency laws, or capacities, that operate robustly across changes in boundary conditions, not just in experimental conditions. This is because the experiment may be so designed that although background factors are held fixed, the outcome depends on their non-additive interaction with the main experimental variables.

What about the external validity of the outcomes of simulations? *Quantitative* conclusions are usually warranted only if model and target system are similar enough, and values of variables are known. This applies to, e.g., complex—but non-chaotic—systems whose interactions with the environment are well captured by the model. And it applies to chaotic systems, too, although to a lesser degree. For instance, on the assumption of the approximate linearity of the local dynamics of chaotic systems, 'tessellation' may provide short-term predictions of chaotic behaviour (see §1.3.2).[38]

Also, contrary to more traditional experiments, even when the conditions for drawing quantitative conclusions cannot be met simulations may still warrant *qualitative* conclusions, based on the exploration of a wider space of possibilities. Problems of subjects sample are here constituted by too few runs, or too little variation in initial conditions. These problems are relatively easy to solve, at least conceptually. Problems having to do with bad design cannot be solved by randomisation: since the design *is* the model, and is unique, there aren't different situations where different mechanisms are at work, so randomisation over initial conditions or parameter values alone won't do. However, the transparency of the design facilitates performing sensitivity analyses. If the model can be interpreted in relation to the target, then simulation, in conjunction with sensitivity analysis, helps establish conclusions as regards the presence of marks[39] and robust capacities, and ameliorates the

[38] Other cases of quantitative conclusions allowed by chaotic models are those that depend on non-universal parameters, such as Liapunov exponents, the fractal dimension of strange attractors, indexes of bifurcation rates (see Smith, 1998, p. 118), or power-law exponents corrected for the presence of log-periodic modulations (Sornette, 2002).

[39] This procedure is called 'benchmarking', and consists in comparing outputs of the simulation and known facts about the target (see Frigg and Reiss, 2009, p. 603).

problem of missing data on the target situation.

For the moment, we need not be concerned with the issue of the validity of the models presented in this chapter. In chapters 4 through 6, I will proceed on the assumption that the models faithfully represent causal relations in the apoptosis and the asset pricing mechanism, that is, they identify salient features that are causally responsible for, respectively, the irreversibility of apoptosis, and volatility and crashes. I will then come back to the bearing of the validity of the models of complex systems in chapter 8, where I illustrate how my inferentialist account addresses the issue of the objectivity of causal claims in complex systems.

Conclusion

I introduced aims and methods of systems biology and computational economics, as well as two phenomena studied by these disciplines, viz. apoptosis and asset pricing. I described models built to account for such phenomena. I defended the possibility of giving a causal interpretation to such models and to the causal claims derived with their aid. My task will now be to provide a suitable interpretation of the meaning of these claims. In particular, I will ask: Which account of causality provides the most adequate analysis of 'causes' as occurring in such claims? And how do these considerations affect our understanding of causality *in general*?

Chapter 4

Difference-making Accounts of Causality

In this chapter I discuss the so-called 'difference-making' accounts of causality. They are based on the intuition that a cause is something that makes a difference to the effect. One can distinguish between *reductive* accounts, such as the counterfactual account (§4.2) and the agency account (§4.4), and *non-reductive* accounts, such as the contextual unanimity account (§4.3) and the interventionist account (§4.4). Difference-making accounts try to elucidate the notion of causality in terms of one or the other privileged criterion, or *test* condition, which typically grants the inference to the causal claim. Such test conditions are then usually erected to either truth conditions, relevant to the obtaining of mind-independent causal relations, or to conceptual analyses, relevant to our understanding of the notion of causality. The appeal of difference-making accounts in complex systems comes from the fact that causes are—typically—difference-makers. Yet, difference-making criteria fail to *exhaustively* capture the meaning of causal claims in complex systems. There is more to 'causes' than difference making.

4.1 Regularity accounts

Regularity accounts are committed to the following two tenets: (i) causal relations are *general*, that is, must be analysed in terms of properties of classes of events, more specifically facts about regularities among events belonging to those classes; (ii) causal relations are *invariant*, regular association being invariable succession between events in such classes.[40]

The regularity view of causation is a direct descendent of Hume's definition: A causes B iff A is spatially contiguous to B, A is temporally prior to B,

[40]This distinction is borrowed from Arif Ahmed. Notice that the two tenets can be held independently. Endorsement of one and rejection of the other result in different (difference-making) views on causality (see §4.2 and §4.3).

and all A-type events are regularly followed by (or constantly conjoined with) B-type events (see Hume, 1968, I.iii.14). The regularist can either maintain that there is nothing more to causation than regularities (strong version) or admit that the notion of causality cannot be fully reduced to regularities (weak version).[41] Either way, he maintains that the notion of causation is best elucidated—whether partially or totally—by reference to the more transparent notion of regularity, viz. a kind of dependence.

The main troubles with regularity accounts are accounting for relations involving imperfect regularities, recovering the asymmetry of causation, and avoiding spurious correlations. Here I evaluate to what extent regularity accounts can cope with such problems and be applied to complex systems with reference to Baumgartner (2008)'s version of the regularity theory (in short, RT), which improves in many respects traditional versions of the regularity account (most notably Mackie, 1974). In RT, A causes B iff:

[**RT**$_1$] A is part of a minimally sufficient condition AX_1 of B;

[**RT**$_2$] AX_1 is a disjunct in a disjunction $AX_1 \vee X_2 \vee ... \vee X_n$ (or "minimal theory" Φ) of other minimally sufficient conditions of B ($n \geq 2$), the disjunction being minimally necessary for B;

[**RT**$_3$] A is part of Φ and stays part of Φ across all extensions of the variable set considered;

[**RT**$_4$] The instances of $AX_1 \vee X_2 \vee ... \vee X_n$ and B differ and are spatiotemporally proximate.

RT purports to distinguish genuine from accidental and spurious regularities, to distinguish causes from effects, and to allow for non-exceptionless regularities to count as causal. I will first illustrate RT and then evaluate to what extent it is successful when applied to causal claims in complex systems.

[41] The weak version is embraced by, e.g., Mackie (1974). Here, two further interpretations of Hume are open to the regularist. First, a skeptical realist reading: there are causal necessities in nature but the secret connexion that underlies them cannot be reduced to regularities (Strawson, 1989). Secondly, a quasi-realist reading: there are no necessary connections in nature (they are the result of a human 'projection'), although there may be objective causal relations *insofar as* there are regular associations (Blackburn, 1990).

As in the Humean regularity account, causes and effects are spatiotemporally distinct yet proximate (RT_4).

Minimal sufficiency (RT_1) is invoked to overcome the inability of the Humean analysis to distinguish between genuine and accidental regularities. Any sufficient set of factors must be non-redundant, i.e. contain no conditions that just happen to be constantly followed by certain events without making a difference to them. Redundant factors do not count as causes.

RT_2 is imposed to deal with non-exceptionless regularities, to distinguish causes from effects, and to distinguish spurious from genuine regularities.

First, not all causes are regularly associated with effects. For instance, DNA damage causes increase in p53, which in turn causes synthesis of proapoptotic proteins. However, p53 increase is not always associated with synthesis of proapoptotic proteins. This may depend, among other things, on whether the gene that codes for p53 is mutated. Similarly, in the stock market the traders' sensitivities to opinion, price changes and profits being below some threshold prevents crashes. However, sensitivities below the threshold are regularly associated with absence of crash only if the proportion of chartists does not exceed a critical value. RT explains why a cause is (or is not) followed by the effect in terms of whether the *other* conjuncts are instantiated. In general, causal relations among two conditions are relativised to dependences in complex sets of conditions. In this way, RT inherits the merits of Mackie's account. For Mackie, not only total causes but also causal *factors*, that is insufficient but non-redundant parts of unnecessary but sufficient (INUS) conditions, count as causes. Non-exceptionless regularities may count as causal, since INUS conditions are not sufficient for the effect.

Secondly, on the assumption that in complex nets of causal factors the cause *overdetermines* the effect but not vice versa, regularities suffice to distinguish causes from effects. This should allow for there being at once *nomological dependence* between cause and effect (since any disjunct in Φ is sufficient for B, the presence of the effect can be inferred provided the presence of some of its causes) and *inferential asymmetry* (since B is only sufficient for the whole disjunction, not for any particular disjunct, the presence of a specific cause cannot be inferred from the presence of the effect).

Thirdly, by imposing *minimal necessity*, too, RT_2 ensures that spurious regularities are excluded via the exclusion of redundant disjuncts. This makes RT superior to Mackie's account. Consider the following structure, where A (the sounding of the hooter in Manchester) and B (the London's workers

leaving from work) are the effects of a common cause C (a certain time of the day), D is a further cause of A, and E is a further cause of B. Since $A\overline{D}$ is minimally sufficient for B and part of a necessary condition for B, such that $(A\overline{D} \vee C \vee E) \leftrightarrow B$, then $A\overline{D}$ would count as a cause of B. RT avoids this by imposing that necessary conditions, too, be minimal. $C \vee E$ is minimally necessary for B (there are no instances of B without either C or E), but $A\overline{D} \vee C$ and $A\overline{D} \vee E$ are not (there are instances of B without instances of $A\overline{D} \vee C$ and $A\overline{D} \vee E$).

Finally, RT_3 guarantees that if some A is a genuine cause of B in an incomplete theory Φ of B (such that, e.g., A and C are sufficient for B, but B is not sufficient for $A \vee C$), then A is not made redundant by the discovery of other conditions for B (i.e., there is no X that is sufficient for B, and such that B is sufficient for $X \vee C$).

However, RT is not well suited to analyse the meaning of causal claims in complex systems. To begin with, one may question the role of the *number* of variables and Φ's extensions. Incomplete knowledge of regularities grants causal inference only by assuming that the cause stays part of Φ across all extensions of Φ. Since there is no way to *actually* test all possible extensions, it seems that the criterion legitimates causal inference only at the price of trivialising the analysis. But let us assume that attempts to test causality against large variable sets are enough for practical purposes. Still, the role of regularities in such attempts does not serve the purpose of conceptual reduction, as RT maintains. This point can be illustrated by reference to the LN in (Mai and Liu, 2009). LNs are a very natural way to encode and study regularities in large variable sets. However, Mai and Liu (2009)'s LN is useful not because it allows one to bootstrap causal relations from pure regularities, but because it helps to study macro-dependences on the previous assumption of causal relations among individual variables, in turn based on literature and expert knowledge. In short: no causes in, no causes out. Also, for the model to provide significant statistics, even assuming a good deal of causal knowledge, a vast amount of data must be available. Since databases from empirical studies are too small to provide a representative sample, the model's conclusion must rely on random samplings of initial conditions and simulations. In other words, the regularities employed to test causal relations are not actual but artificially produced. This is not to say that the model is useless. However, the larger the variable set, the less informative the reduction of causation to observed regularities.

But there is more. In RT, the possibility of distinguishing genuine from accidental and spurious regularities, and causes from effects, depends on the existence of complex structures of regularities. Given any two factors, the complexity of the context—that is, the regularities amongst *many* factors—should determine whether a causal relation obtains between them, and if so, which factor is the cause and which is the effect. However, the number of variables is not the only determinant of complexity. Crucial are also the *nature* of the variables and their mutual *relations*. Such features make RT inadequate in complex systems.

For instance, chaotic behaviour (e.g., Lorenz system, the markets in §3.2.3) depends on the way a few non-linear interactions result in a strange attractor. If the system is chaotic, the relata may not be subsumable under a regularity. Two are the possibilities: either the relata are specific states of the system, or one (the cause) is a set of parameter values and the other (the effect) is a class of similar events, or motifs. If the relata are states, any two same-time states, however close, lie on diverging trajectories, so 'like cause, like effect' claims are false: the less proximate the effects, the greater the difference. If the cause is any of the sets of parameter values determining chaos, this is responsible for *all* sorts of motifs, and there is no clear sense in which such sets are regularly associated with the motifs. Furthermore, in chaotic systems RT cannot recover the asymmetry of causation from the inferential asymmetries between causes and effects. In fact, here the complexity depends not on the number of variables—in particular, the existence of several sufficient causes—but on the sensitive dependence to initial conditions that result from the (non-linear) way the variables are related to one another. Even leaving chaos aside, there are further problems with RT.

In general, RT should be able to deal with imperfect dependences due to the *non-linear interactions* among the variables, common to complex systems. However, non-linearities often result in the absence of regularities. Even assuming that genuine causal relations between variables can be identified, and that non-linearities don't give rise to chaotic behaviour, non-linearities can make the *direction* of the effect sensitive to the values of the other variables (see Wagner, 1999, p. 95). In short, the larger the number of non-linear interactions, the larger the number of possible attractors, and the more sensitive the regularities to variations in context. And the greater this kind of context sensitivity, the less applicable the 'like cause, like effect' principle. In fact, for the direction of change to be specified in terms of regularities, results of

changes in one variable must be specified with respect to all possible combinations of values of all variables. Assume one starts with no model of the system, only with data, and tries to discover causal relations by identifying regularities. If each causal relation depends on the specification of regular changes, the regularities may be so fine-grained that they will be judged either causal *by fiat*, because trivially exceptionless, or non-causal, since any sample is too small to be representative.

Last but not least is the issue of *determinism*. RT works on the default assumption of universal determinism, which entails that for each effect there is always at least one set of sufficient conditions for it. Although there may be instances where it is appropriate to conceive of causes as sufficient (or necessitating) conditions (see §2.1.3), there may *not always* be such conditions, so one should not rely on their existence to analyse causality. Since complex systems phenomena may be the result of—at least partially—indeterministic processes (e.g., superconductivity, ferromagnetism), one should better not build determinism in the analysis of causality, on pain of making it unnecessarily restrictive.

In the light of these problems, one may attempt to modify the difference-making analysis by abandoning either one or the other tenet to which regularity accounts are committed, viz. generality (§4.2) and invariance (§4.3).

4.2 Counterfactual accounts

Need causal claims be general? Why should the truth of a 'this causes that' claim depend on what happens at other places and times? It might well be the case that there are causal claims which are generally true, e.g., strict laws. Yet, one may argue that this fact depends on the *local* conditions which link cause and effect *in each instance* of the regularity. Based on this 'singularist' intuition,[42] the counterfactual account abandons the requirement of generality whilst maintaining invariance.

Besides the regularity definition, Hume offers also a second, different definition: C causes E iff "if the first object had not been, the second never had existed" (Hume, 1975, §7.ii.60).[43] The most renowned development of this idea is Lewis' counterfactual dependence account of causation—only for

[42] This intuition is also common among mechanistic accounts, for which it is the productive nature of what happens between the relata that makes a relation causal (see chapter 5).

[43] Interestingly, Hume failed to notice the difference and took the two definitions to be equivalent.

Lewis the relata are *events* rather than objects.

An event, for Lewis, is a property, or class, of spatiotemporal regions. Properties of regions that are genuine events are *intrinsic*, in the sense that they 'supervene on' the region alone, and do not depend on extrinsic features of the world. Causation supervenes on the totality of events (intrinsically considered) and fundamental matters of fact (whether deterministic or indeterministic) holding at a world. Causation is defined as 'transitive counterfactual dependence', in two steps.

First, Lewis characterises 'causation' in terms of 'causal dependence'—of which causation is the 'ancestral'. Causation between events is defined as *transitive* causal dependence along a *chain* of events:

> event c causes event e iff either e depends on c, or e depends on an intermediate event d which in turn depends on c, or... (Lewis, 1986, p. 242).

That is, causation between c and e depends on whether the causal dependence between c and e is either direct or mediated by a chain of events such that each pair of successive events in the chain stand in a relation of causal dependence. For Lewis, transitivity is necessary for causation. This is to cope with cases of symmetric preemption. Here is an example: A and B are both sufficient for C; however if A operates, B does not, and vice versa; in the actual case A operates, thereby preempting B and causing C; however, C does not causally depend on A. Lewis solves this problem by requiring that there be a chain of events between the cause and the effect. Since C transitively depends on A, and not B, A is the cause.

Then, Lewis reduces (direct) 'causal dependence' to 'counterfactual dependence':

> Causal dependence is counterfactual dependence between distinct events. Event e depends causally on the distinct event c iff, if c had not occurred, e would not have occurred—or, at any rate, e's chance of occurring would have been very much less than it actually was (Lewis, 1986, p. 242).

The counterfactual dependence between events in implication- or mereological-relation with one another is excluded as non-causal by imposing that the events in the causal relation be distinct, i.e., they cannot be properties of

the same region or of overlapping regions (see Lewis, 1986, pp. 256, 259). The clause about chance is meant to help the account to deal with non-deterministic, or chancy, causal relations.[44] In sum, the counterfactual dependence (CD) account reads:

> **[CD]** C causes E iff there is transitive counterfactual dependence between C and E.

Counterfactual dependence holds, strictly speaking, between families of incompatible *propositions* describing relations between possible events. The truth conditions for such propositions are defined over a possible-worlds semantics (Lewis, 1986, pp. 163-166), such that, for any two propositions C and E describing events c and e, '$C \mathbin{\square\!\!\rightarrow} E$' (i.e., '$e$ depends counterfactually on c') is true iff some c-world where e holds is closer to the actual world than is any c-world where e does not hold. Notice that the left-hand-side must hold non-vacuously, that is, c-worlds must exist. Whether a given counterfactual is true in the actual world depends on similarity relations between the actual world and other possible worlds, so that 'had c not been, e would not have been' must be true in all closest worlds to the actual one. Possible worlds and their ordering with respect to their similarity are objective and mind-independent. We have 'access' to such similarity relations by means of a judgement of comparative similarity, based on the following criterion: the more similar are the laws in a possible world to those holding in the actual world, and the more similar is the antecedent complete state of a possible world to the antecedent state of the actual world, the closer is such a possible world to the actual world. For Lewis, the similarity relation is "primitive" (i.e., it cannot be further analysed) and vague, yet "familiar" to all of us (Lewis, 1986, p. 163). In the case of complex systems, a way to judge similarity/closeness of worlds is by reference to models which specify the distance between any pair of possible events at any given time (in all worlds where laws are fixed and the same as in the actual world) in terms of, e.g., distance of points that represent those events in the phase space.

The counterfactual account can be illustrated by means of a simple causal structure, a 'fork' of three events, namely a common cause P and its effects B and R. The classical example, to which I'll refer for ease of exposition,

[44] For more on Lewis' treatment of chancy causation, see Lewis (1986, pp. 175-184).

involves the relation between pressure (P), barometer reading (B) and rain (R). However, it is easy to make the example more relevant to complex systems, by thinking of P as measuring p53 expression, and B and R as two effects of the regulatory activity of p53, e.g. apoptosis and DNA repair (see §3.1.2), measured respectively in terms of levels of caspases (Casp3 or Casp9) and DNA polymerases (Polδ or Polϵ), and neither one causing the other.

So, let P_1 be the proposition 'pressure falls at t', B_1 be the proposition 'the barometer says "rain" at t'', and R_1 be the proposition 'it rains at t'''. And let P_2 be the proposition 'pressure rises at t', B_2 be the proposition 'the barometer says "fine" at t'', and R_2 the proposition 'it does not rain at t'''. And let p_1, p_2, b_1, b_2, r_1, r_2 stand for the corresponding events. According to the counterfactual account: (i) P causes B and R, since the counterfactuals '$P_1 \Box\!\!\rightarrow B_1$' and '$P_2 \Box\!\!\rightarrow B_2$', and '$P_1 \Box\!\!\rightarrow R_1$' and '$P_2 \Box\!\!\rightarrow R_2$', are true; but (ii) B does *not* cause R (nor does R cause B), since the counterfactuals '$B_1 \Box\!\!\rightarrow R_1$' and '$B_2 \Box\!\!\rightarrow R_2$' (respectively, '$R_1 \Box\!\!\rightarrow B_1$' and '$R_2 \Box\!\!\rightarrow B_2$') are false.[45] That is, the relation between barometer reading and rain is ruled out as spurious, since the counterfactuals 'hadn't the barometer said "rain", it wouldn't have rained' and 'hadn't the barometer said "fine", it would have rained' are false. In fact, one can imagine that the closest possible world to the actual one is one where all matters of fact—including those relating pressure to rain—are the same, with the only exception that the barometer is broken or biased. In that case, the counterfactuals are false, hence the corresponding causal dependence relation between barometer reading and rain does not obtain.

How does the counterfactual account get the temporal asymmetry of causal dependence right? Lewis' idea is that there always are counterfactual asymmetries between antecedent and consequent states of causal relations. It is a *contingent* fact that at any time the set of events which are jointly sufficient for the occurrence of some *later* event is smaller than the set of the events which are jointly sufficient for the occurrence of some *earlier* event. The former overdetermine the latter, not vice versa. Accordingly, only a *small* and *local* miracle in the past light cone of an event is sufficient for the non-occurrence of the event to be counterfactually dependent on such a miracle.

[45] Notice that Lewis does not allow 'backtracking' reasoning to enter the judgement on causal relations. This means that we are not allowed to consider counterfactuals such as 'if b_1 had not occurred, it would have to *have been* the case that p_1 did not occur, in which case r_1 would not have occurred'. Only non-backtracking counterfactuals may be employed, i.e. counterfactuals that hold the past fixed up until the counterfactual antecedent event. Lewis proposes that non-backtracking counterfactuals are distinguishable from backtracking counterfactuals based on their greater similarity to actuality.

Instead, many, non-local miracles in the future light cone of the event would be needed for the event to counterfactually depend on them. For instance, take the event of rain in a given place at a given time. Only a local change in the pressure at some previous time in the vicinity of that region—viz. the *cause* of rain—is sufficient, so the story goes, to counterfactually affect rain. Instead, many later, non-local events such as maintenance of the local ecosystem, production of hydroelectric power, crop irrigation, etc.—viz. the *effects* of rain—would be needed to make a counterfactual difference to rain.

Does Lewis' counterfactual analysis succeed in capturing the meaning of causal claims in complex systems? Not completely. To begin with, one may object that Lewis' possible-worlds semantics is too vague and ambiguous for the notion of counterfactual dependence to really illuminate the concept of causality. How can one decide what counts as a 'small' miracle? And, since we have no contact with possible worlds, how do we know what world is closest to actuality? For instance, it is well-known that our causal judgments are sensitive to the way the events are individuated. We will judge the truth of 'The camper's lighting of the fire caused the forest's destruction' differently, depending on whether we contrast the antecedent event with one where the camper doesn't light the fire at all, or one where the camper lights the fire in a slightly different manner or at a slightly different place or time. (Lewis, 2004)' most recent view is that the problem depends on linguistic indeterminacy and there is no principled way to solve it. However, complex systems show that even with respect to models where the possible-worlds semantics *can* be interpreted unambiguously the counterfactual account is inadequate.

Consider models of chaotic systems, for which the notions of similarity and miracles *can* be made precise. In analogy with the distinction introduced in §4.1, one can distinguish between two possible interpretations of c, as either a fragile state or sequence of states (a motif), or as a non-fragile set of parameter values. In the former case, where the chaotic model is taken to represent *literally*, there is no possible c-world besides the actual world. Here the counterfactual is true (the closest non-c world is a non-e world), but trivially so (all possible worlds are non-c). In this case, the criterion is too strict to be informative.[46] The latter case is even more interesting, as it leads

[46] Another case are chaotic systems where the sensitive dependence depends on initial conditions belonging to a fractal basin of attraction (see fn. 7). Here, although there are c-worlds besides the actual one, if the model is interpreted literally, i.e., there is a physical state corresponding to each point in the phase space, it is conceptually impossible to try to find the closest non-c world: between any two initial conditions—however close—ending up in the same final state, there is another one leading up to a different final state. So, it

to neither triviality nor undecidability, but rather shows that causality need not entail counterfactual dependence. If we take c to be a set of parameter values, the model can be interpreted as representing more *loosely*: the exact details and timing of the states of the system are irrelevant. Then, it is *false* that the closest non-c world is a non-e world. Take Lux and Marchesi's market (§3.2.3) and consider the actual world to be one where sensitivities above the threshold cause a crash. Since the occurrence of crashes is robust across changes in parameter values, the closest non-c world (where the parameters are slightly different, e.g., one or more are below the critical value) is such that sooner or later the proportion of chartists will exceed the critical value and cause a crash. So, there can be causation and no counterfactual dependence. Either way, no matter what our intuitions are, CD is inadequate.

Nor need one rely on the special case of chaotic systems to come to these conclusions. There are, in fact, other cases such that there is causation but either (i) the counterfactual account does not say what causes what or (ii) there is no transitive counterfactual dependence. To the first class belong cases of symmetric overdetermination. For instance, p53 and TNF are (*ceteris paribus*) individually sufficient to affect Casp3. In the circumstances, they act together. Does Casp3 counterfactually depend on both of them, or only one of them? It is not clear how the account can deliver a verdict on what causes what.[47] To the second class belong cases of non-symmetrical overdetermination (e.g., chancy preemption), where there is no dependence, and cases where causes and/or effects are absences (e.g., double prevention), where there is no transitivity.

I illustrate non-symmetrical overdetermination with reference to a chancy preemption case in complex—but non-chaotic—systems. Chancy preemption is such that there are two processes, one preempting the other, which lead to the effect with different probabilities; in the circumstances, the low-probability process preempts the high-probability process but, nevertheless, produces the effect. As a result, the preempting cause *is* the actual cause even though the effect does *not* counterfactually depend on it, and the preempted cause is *not* the actual cause even though the effect *does* counterfactually depend on it. To illustrate the chancy preemption case, one may consider

is *in principle* undecidable whether the closest non-c world is an e or a non-e world.

[47] Notice the use of the *ceteris paribus* clause, which signals that the causal dependence crucially depends not only on putative cause and effect and inbetween events but also on the presence of suitable boundary conditions. I discuss the role of *ceteris paribus* conditions below, when I show that causal relations are contextual, not intrinsic—against Lewis' intuition.

the study of the LN network, where perturbations on internal nodes are performed to study the changes in the chance of an otherwise surviving state, that is, whether the survival-to-apoptosis transition becomes more or less likely. In the absence of the TNF signal, GF_{ON} would lead to survival with high probability. Setting GF to OFF makes survival less stable, i.e., it raises the probability of survival-to-apoptosis transitions. But as it happens, GF_{OFF} ensues in a configuration that, although preempting the higher-probability process for survival, still results in survival. Here, survival does not counterfactually depend on the preempting lower-probability process, which is the actual cause; instead, it counterfactually depends on the preempted higher-probability process, which is not the actual cause.

Let me turn to counterexamples involving absences. As mentioned, Lewis requires transitivity to avoid symmetric preemption. The problem is that in other cases—e.g., late preemption and double prevention—causal dependences involve no chain of events, that is, they are not transitive.

In the late preemption case, both X and Y are sufficient to cause Z. Y is such that it operates only if X does not. In the circumstances, X operates, preempts the later process Y, and causes Z. However, Z does not counterfactually depend on X. The counterfactualist might deal with this problem by appealing to the *intrinsic resemblance* of the actual scenario to another scenario where there is transitive counterfactual dependence between cause and effect and the preempted alternative is absent.[48] The idea is that only events, i.e. essential properties of specific spatiotemporal regions, can be causally efficacious. The absence of Y is not a genuine event, but a disjunctive fact: either this event is instantiated, or that event is, or... For Lewis (1986, pp. 172-173), although there is nothing in the XYZ region which makes X be the cause of Z, X still qualifies as the cause, in virtue of its intrinsic resemblance to other X-events that cause Z in the absence of preempted causes Y. Actual processes that don't exhibit counterfactual dependence are 'causal by courtesy', i.e., they exhibit 'quasi-dependence', in virtue of their intrinsic resemblance to the comparison case.

[48] Lewis (1986, Postscripts to 'Causation') and Menzies (1996) go this way. In particular, Menzies offers a 'folk theory' of causation, where causation is defined as a theoretical entity in a theory of platitudes about the concept of causation: causation is an intrinsic relation between events; it is 'typically' accompanied by counterfactual dependence; etc. Here, causation is not defined *reductively* as such-and-such a relation, but *functionally*, as whatever worldly relation occupies the role the theoretical entity has in the theory. This strategy, however, is unsuccessful because the modified account does not apply to all cases of causation, e.g., double prevention (see below).

However, the appeal to intrinsic resemblance does not help elucidate the meaning of causal claims in complex systems. To begin with, the resulting account goes against the intuition that causation can involve omissions/absences and the common practice of scientists of explaining by ascribing causal role to absences and disjunctive kinds.[49] One might think: too bad for the intuition. But the account has further problems. First, since the account violates the requirement that the actual world is one where the counterfactual dependence is instantiated, it fails to constitute a reduction of causation to counterfactual dependence. Secondly, and more importantly, the account doesn't apply to double prevention cases.

In double prevention cases, contrary to late preemption cases, there is no connecting process between cause and effect, so transitivity fails. For instance, Apaf1 and XIAP both cause Casp9—more precisely, Apaf1 *promotes* Casp9, whereas XIAP *inhibits* Casp9—which then causes Casp3. Casp3, in turn, promotes its own activation, for instance by preventing XIAP from preventing Casp9 to cause Casp3. Casp3's high level causes more Casp3 activation. However, there is no connecting process between Casp3 binding to XIAP and Casp9 promoting Casp3—part of the spatiotemporal region between the two events is not occupied by a chain of events that serves as a connecting process. Here the relation is causal not in virtue of the transitivity of the counterfactual dependence, but in virtue of the contextual features that make it possible for the effect to still causally depend on the cause, even in the absence of a causal chain. In general, double prevention cases show that causal responsibility is attributed not on the basis of their intrinsic resemblance to cases where there is transitive counterfactual dependence, but rather on the basis of the *context*, viz. features *extrinsic* to the putative cause and effect and what is between them. The *ceteris paribus* clause is meant to describe exactly such contextual features. This shows that causation need not entail transitive counterfactual dependence. And since the sort of context sensitivity typical of double prevention cases is widespread in complex systems, it is all the more implausible to analyse the meaning of causal claims in complex systems in terms of (their intrinsic resemblance to) relations of transitive counterfactual dependence.

Finally, it is not true that counterfactual dependence is sufficient for causation. This can be shown by considering how the counterfactual account

[49] Structural equation accounts (see Menzies (2009, §4.2) and Hitchcock (2010, §4.1)) attempt to solve this problem whilst trying to remain faithful to the intuition that causation is counterfactual dependence. More on this in §4.4.

handles the asymmetry of causation in terms of asymmetry of miracles. There are two classes of cases the account must deal with, depending on whether the laws are deterministic or not. If laws are deterministic, since they are stated in terms of necessary and sufficient conditions, they are symmetric. In this case, it is as true that a small change in the consequent state determines a change in the antecedent state as is true that a small change in the antecedent state determines a change in the consequent state. There is no way to tell which is which unless (perhaps) by considering the locality of antecedent and consequent states. For Lewis, in fact, prior determinants are localised, whereas later determinants are not. But not always this criterion can be employed, for instance if the laws are non-deterministic. Here, the usual counterfactualist move is to claim that facts about entropy increase make it the case that later determinants are more sensitive to prior determinants than vice versa. However, it is controversial whether this reply is successful (cf. Elga, 2000). At least in self-organising systems such as Bènard rolls, where many coordinated events result in some emergent phenomenon, the dependence of the emergent phenomenon on the microstates is robust, or non-sensitive: a change in one of the many prior determinants does not result in a major change in the effect, only in a change in the effect's exact magnitude and/or timing, etc. The markets in §3.2.3 count as self-organising in this sense. In self-organising systems, it is false that the miracle that changes the effect (e.g., a crash) by changing its prior determinant(s) (e.g., a change in the individual behaviour of the traders) is smaller than the miracle that changes the cause by changing its later determinant(s) (e.g., a forced suspension of trading due to the excessive drop in asset prices). Also, the change in the later determinant may be as local as, or more local than, the change in the prior determinants. Here we do have counterfactual dependence—in the effect-to-cause direction—and yet no causation—since the asymmetry goes in the wrong direction.

4.3 Probabilistic accounts

Need causal relations hold invariantly, as in regularity and counterfactual accounts? Let us consider cases where we have positive evidence of genuinely indeterministic causation, e.g., radioactive decay. In these cases, our best theories suggest that no further knowledge can help determine whether or not a given phenomenon will necessarily obtain after another. What does

causality mean in these contexts? An obvious alternative is to analyse causal claims in terms of probabilities. Probabilistic accounts envisage causal claims as *non-invariant*, contrary to regularity and counterfactual accounts, and—usually, but not necessarily—as *general*.[50] In probabilistic accounts, a cause is an event type, or property, that makes a probabilistic difference to another event type, or property, viz. the effect. In general, probabilistic analyses consist of two steps: first, a characterisation or analysis of the notion of causation in terms of probability, and then, an explication of probability in terms of, e.g., relative frequencies, degrees of belief, etc. I will here assume that the second step has been dealt with and focus on the first step.

Probabilistic accounts can be categorised into (i) accounts based on the intuition that a cause makes a positive difference to the effect (probability-raising accounts), and (ii) accounts that do not require this. In both categories one finds both conceptually reductive accounts and non-conceptually reductive accounts.

Let us indicate with uppercase letters variables and with lowercase letters values of variables, so that C and E stand for, respectively, the cause variable and the effect variable, and c_1 and c_0 (respectively, e_1 and e_0) stand for, respectively, the obtaining and the non-obtaining of the cause (the effect). (Here, C and E are binary variables.) All probability-raising accounts assume the following, probability-raising condition (PR):

[PR] C causes E iff $P(e_1|c_1) > P(e_1|c_0)$.

PR alone, however, is not sufficient to analyse causation, because it does not deal with asymmetries and spurious correlations. First, probability raising is, in itself, symmetric: $P(e_1|c_1) > P(e_1|c_0)$ iff $P(c_1|e_1) > P(c_1|e_0)$. Thus, PR alone does not determine whether C is the cause of E or vice versa. Secondly, take the case where B (level of mercury in a barometer) and R (occurrence of rain) are both caused by P (pressure). In that case, it may be that $P(r_1|b_1) > P(r_1|b_0)$ even if B does not cause R. In the light of these considerations, two main strategies have been pursued to salvage the probabilistic account. The

[50] A *single-case*—as opposed to a generalist—probabilistic analysis of causality may require an intensional, counterfactual semantics to analyse the notion of propensity (see, e.g., Fetzer, 1970, 1981). Alternatively, one may have a single-case analysis in terms of rational degrees of belief. *Generalist* analyses, instead, are usually based on an extensional semantics and a frequency (whether long-run or limiting frequency) interpretation of probability.

first is to merely *characterise* causality in terms of probability raising, rather than reduce the former to the latter (§4.3.1). The second is to focus on the notion of *probabilistic difference* and try to build a probabilistic account—whether conceptually reductive or not—around this notion (§4.3.2). Both strategies face problems, and neither applies well to complex systems. In both cases, the failures have led, for quite separate reasons, to modify the probabilistic analysis so as to account for the meaning of causal claims not in terms of probabilities alone, but in terms of the probabilities generated by the result of interventions (see §4.4).

4.3.1 Probability-raising accounts

Let me start with modifications to PR. Cartwright (1979) offers an analysis of the mutual constraints of causal laws and laws of association (CC):

> **[CC]** C causes E iff c_1 increases the probability of e_1 in every situation otherwise causally homogeneous with respect to E.

CC is a non-reductive analysis of 'causes', because causation isn't reduced to probability raising alone (to recover causal asymmetry, temporal ordering is assumed), and the word "causal" appears in the analysis itself (although not to characterise the relation between the two variables of interest). Yet, at the level of specific causal claims CC provides a sort of conceptual reduction: given temporal ordering and a causally homogeneous background, all there is to causation between two target variables is probability raising. A causally homogeneous background is described as one of the possible conjunctions $k_i = \wedge \pm c_i$ of factors causally relevant to E. CC says that, C causes E iff c_1 is positively relevant to the probability of e_1 across all possible combinations of such factors (excluding C and any intermediate causal factor between C and E).[51] For instance, CC says that holding fixed the context, but not the values of p53 and its downstream effects, if (high, or wild) p53 always increases the probability of apoptosis, then p53 causes apoptosis. However, consideration of causes with *dual capacities*, which operate along different paths, and of *interacting capacities*, which produce their effect depending on

[51] For the emphasis on positive probabilistic relevance in all causally relevant contexts, this criterion is usually referred to as 'Contextual Unanimity Theory'.

their mutual interactions or interactions with the context, led Cartwright to abandon CC.

For instance, p53 has a dual capacity: it can *promote* apoptosis, by promoting synthesis of proapoptotic proteins, and *prevent* apoptosis, by promoting DNA repair.[52] How could one determine whether or not p53 is positively relevant to apoptosis across all contexts, thereby being a cause of apoptosis? Holding fixed the context at the time p53 level is measured and then checking that $\forall i\, P(\text{apop}|\text{p53} \wedge k_i) > P(\text{apop}|k_i)$ won't do. In fact, one cannot distinguish whether an increase in apoptosis obtains *in virtue of* p53 triggering synthesis of proapoptotic proteins or *in spite of* p53 activating DNA repair. What one should do is determine whether p53's effect on the death of cells which *wouldn't have died* were it not for p53 is—always—greater than p53's effect on the death of cells which *wouldn't have been repaired* were it not for p53 (cf. Cartwright, 1989, p. 101).

Usually, one tackles the problem by holding fixed the value of intermediate causal factors along all paths between C and E except the path where one wants to calculate C's relevance to E, for all k_i. For instance, assuming the existence of only two paths from p53 to apoptosis, one via Casp9 and the other via Polδ, to check the relevance of p53 along the Casp9 path one first holds fixed Polδ at the time p53's effect on apoptosis is measured and then checks whether $\forall i\, P(\text{apop}|\text{p53} \wedge k_i \wedge \text{Pol}\delta) > P(\text{apop}|k_i \wedge \text{Pol}\delta)$.

However, there may be contexts such that the influences of p53 and k_i on Casp9 are not independent of one another, in which case what matters is their joint effect, and not that of p53 alone. For instance, p53 and k_i may (nonlinearly) interact with one another so that their joint effect on apoptosis is positively relevant, although conditional on Polδ and k_i, p53 is negatively relevant. Yet, it seems legitimate to claim that p53 is *causally relevant* to apoptosis even if it does not raise its probability across all contexts. Importantly, such interactions are very common in complex systems, where the probabilistic relevance of causes on effects is often sensitive to the context, due to the presence of many factors and nonlinear interactions among them, as evidenced by the apoptosis and asset pricing cases. So, CC is not necessary to causation in complex systems.

[52] The original example discussed by Cartwright (1989) involves the dual capacity of contraceptives to prevent thrombosis, via the prevention of pregnancy, and promote thrombosis, via the release of a harmful chemical in the bloodstream (Hesslow, 1976).

Causal relations are *population relative*, or *context sensitive*: a factor which makes a positive difference to another in one case may make a negative difference, or be neutral, in another case.[53] On the face of this, one may still want to maintain that a factor that can make a probabilistic difference counts as 'causal' across all such contexts, viz. it is a 'promoter' in certain contexts and a 'preventative' in other contexts. Then, the question is: *In virtue of what* is the factor causal across contexts, even though it does not always raise the probability of the effect?

A way to go is to say that for something to count as a cause it is enough that it raises the *overall* probability of the effect. In line with this interpretation is the *average causal effect* criterion for causality, which is a weaker criterion than CC: the positive difference the cause makes in some populations outruns the negative difference it makes in other populations. According to this interpretation, the positive difference will manifest itself if the putative cause is appropriately manipulated, for instance by a randomised control trial (RCT). An RCT establishes that a factor is causal if, after randomisation and partition of the population into two subpopulations, an intervention on the putative cause in the test population makes a difference to the value of the putative effect, such that the values of the effect variable in the control population and the test population are different. Since in this interpretation the meaning of causal claims depends not so much on *observed* probabilistic differences, but on results of *interventions*, I postpone discussion of this proposal to §4.4, where manipulability accounts are presented.

An alternative way to fix the probabilistic account is to regard something as a cause if it makes a probabilistic difference (whether positive or negative) to the effect, with respect to a *causal model*, which is often conceived as a Bayesian network causally interpreted.

4.3.2 Bayesian-networks accounts

A causal model consists of a set V of variables and two mathematical structures defined over V, namely a DAG and a probability distribution satisfying the Markov Condition (Pearl, 2000; Spirtes et al., 1993). A DAG is a directed acyclic graph, i.e. a set of directed edges among variables in V such that there

[53] For Cartwright, this means that singular causal facts—facts about the putative cause's "capacities" with respect to the putative effect—are primary, and cannot be reduced to generic ones if we want to pick out the right regularities at the general level. More on this in §5.5.1.

are no loops, i.e., it is not possible to start from a vertice and, by following a path along the directed edges, come back to it. A variable X in the graph is the 'parent' of another variable Y just in case there is an arrow from X to Y, and is an 'ancestor' of Y (and Y is a 'descendant' of X) just in case there is a 'directed path' from X to Y. Associated to the graph is a probability distribution that satisfies the Markov Condition (MC):

[MC] For any X in V and every set Y of variables in V\DE(X), $P(X|\text{PA}(X)\&\text{Y}) = P(X|\text{PA}(X))$

where DE(X) is the set of descendants of X and PA(X) is the set of parents of X. The condition reads: for any variable, the probability of the variable given its parents is independent of the set of its non-descendants. When DAGs are associated with probability distributions that satisfy MC they are called Bayesian Networks (BNs). BNs are a special kind of graph whose nodes represent variables in the domain of interest, and whose arrows represent probabilistic dependences and independences among the variables. When the variables are causally interpreted, MC is called Causal Markov Condition (CMC) and BNs are called causal Bayesian Networks (CBNs).[54] CMC generalises Reichenbach's principle of the common cause (PCC) (see Reichenbach, 1956, p. 163): If A and B are probabilistically dependent, so as to satisfy the relation $P(a_1b_1) > P(a_1)P(b_1)$, either A causes B, or B causes A, or there exists a common cause C such that $P(a_1|c_1) > P(a_1|c_0)$, $P(b_1|c_1) > P(b_1|c_0)$, and the correlation between A and B is 'screened off', that is, A and B are probabilistically independent given C (more formally: $A \perp\!\!\!\perp B|C$). CMC implies the following version of PCC:

[54] BNs can be interpreted as either providing a probabilistic characterisation of causality or as providing a reduction of the notion of causality to the notion of probability. Some (e.g., Spohn, 2002; Pearl, 1988) take BNs to provide an analysis, causal relationships being just the charts of the independences satisfied by probability distributions which meet MC. Instead, others (e.g., Williamson, 2005) regard BNs as a useful way to represent a causal structure, such that many—but not necessarily all—of its relations can be inferred from and in turn produce, probabilistic dependences and independences.

> **[PCC]** If variables A and B are probabilistically dependent, either one causes the other or there is a set U of common causes in V that screens off A and B, i.e., $A \perp\!\!\!\perp B|\mathrm{U}$.

PCC involves an inference from probabilities to causality, and is meant to ensure that BNs can deal with spurious correlations: when a probabilistic relation is observed, we can infer to the presence of a causal relation.

One condition that is usually imposed is that the variables be not distinct for logical or semantic reasons (e.g., mean and variance of the same quantity, or 'bachelor' and 'unmarried man'), in which case their dependence is obviously non-causal. For instance, consider *mixing* of populations with different traits, e.g., a population of cancerous cells, which respond to GF by growing and dividing in an uncontrolled fashion (g_1) and in which p53 phosphorylation does not result in apoptosis (a_0), and a population of healthy cells with opposite features. In the mixed population, growth and division (G) is correlated with apoptosis (A): $P(g_1|a_0) > P(g_1|a_1)$. However, neither one causes the other: G and A are the outcomes of different pathways. And although conditioning on whether cells are healthy or cancerous (H) *does* screen off G from A, it cannot be interpreted as the common cause of G and A: H is partly *defined* in terms of G and A, it does not cause G and A. In this case, arguably the correlation is screened off by variables that aren't logically or semantically related to G and A.

There are several problems with the BN approach. To begin with, not all causal relations may be represented by DAGs, for instance homeostatic mechanisms based on loops of interlocking positive and negative relations. One example is p53 self-regulation, based on p53 promoting Mdm2, which then inhibits p53. Intuitively, the two relations take place at the same time and are both causal. However, DAGs are so designed that the formalism cannot, as it is, account for them.[55]

[55] Various solutions are available to model mechanisms involving loops (Casini et al., 2011, fn. 14). One strategy is to leave out a node (e.g., Mdm2), provided the node in question is not a common cause of other variables. Alternatively, one may combine the values (e.g., a_1, a_2, b_1, b_2) of two variables (A and B) that are connected by a causal loop, into a single variable (AB, taking the possible values a_1b_1, a_1b_2, a_2b_1, a_2b_2). Either way, the relations in the loop are *not* represented by the DAG. Another strategy is to time index the variables, as in dynamic BNs (DBNs) (see, e.g., Friedman et al., 2000). This allows one to say that A causes B and vice versa, however *not at the same time*.

A further issue is that PCC delivers correct conclusions only on the assumption that the set of common causes is *complete*, otherwise common causes may fail to screen off their effects. Consider the case where *two* common causes are present, for instance GF and p53 (respectively C and D) causing cancer (A), by promoting or inhibiting uncontrolled cell division, and cell survival (B), by influencing, whether negatively or positively, apoptosis. However, only C is known. With respect to this causal structure, it may be that $P(a_1b_1|c_1) > P(a_1|c_1)P(b_1|c_1)$, that is, C fails to screen off A and B, due to the residual effect of D. In such a case, PCC simply delivers the wrong conclusion, viz. that C is *not* a cause of A and B. This is a problem whenever it is implausible to assume knowledge of *all* common causes, which is often the case in complex systems, where many causes may be in operation.

Furthermore, sometimes looking for common causes is simply the wrong thing to do. Not all probabilistic dependences are underpinned by relations that satisfy PCC. For instance, consider the case of time series with the same trend (e.g., Venetian sea levels and prices of bread in Britain). They *are* (positively) correlated (increase of one is correlated with increase of the other), and yet neither one causes the other nor do they have a common cause (Reiss, 2007). They just happen to be monotonically increasing time series which are influenced by totally different causes. This is not the case of the relations obtaining in the systems described by the models in chapter 3, simply because each of them is meant to describe the operations of *one* mechanism (in statistical parlance, of one 'data generating process'), whose quantities' correlations are, by assumption, non-spurious.[56] However, PCC fails when applied to series of events which are not the result of the same mechanism, e.g., the series of numbers of cell death events in a biological organism and of crashes in the market. Both series are monotonically increasing, but their correlation is non-causal, since they are the result of independent mechanisms. Importantly, understanding when PCC applies and when it doesn't presupposes *causal* knowledge, that is, the very same knowledge that PCC is meant to deliver.

As I said, CMC involves an inference from probabilities to causality. Another crucial assumption of the BNs approach—viz. *causal faithfulness*—involves a inverse inference, from causality to probabilities: whenever the DAG correctly identifies causal relations, the operation of the system the DAG

[56] As Cartwright (2004) would have it, such models represent the operation of 'regimented systems', responsible for the repeated instantiation of certain events, or states, whose correlations *can* be causally explained.

represents generates probabilistic dependences and independence as represented by the DAG. Otherwise, probabilistic dependences could not be taken as representative of the underlying causal structure, and one could not use CMC to infer causality. However, causal faithfulness is not always met, for instance when causes with dual capacities have opposing tendencies that cancel out. Consider again the case of p53's dual capacity to promote apoptosis via the synthesis of proapoptotic proteins, and to prevent apoptosis via DNA repair. If the capacities along the two pathways cancel out, there may be no probabilistic difference, whether positive or negative, on apoptosis. Usually, the advocates of the BNs approach reply to objections involving violations of faithfulness by appealing to the role of interventions to produce the appropriate probabilistic differences. Since this modified probabilistic account relies crucially on the notion of intervention, I discuss it in §4.4.

4.4 Manipulability accounts

Manipulability accounts are nowadays the most popular amongst difference-making accounts. They try to exploit the conceptual connection between 'causing' something and 'manipulating' something. Their appeal in complex systems comes from the fact that causal claims in complex systems are often accompanied by claims to the point that the effect can be manipulated by intervening on the cause and/or some control parameter, when some other conditions are met (e.g., holding the background context fixed, or eliminating confounding factors).

Manipulability accounts can be categorised into agency accounts, which aim to conceptually *reduce* the notion of causality to that of manipulations, and interventionist accounts, which only aim at *characterising* causality in terms of manipulations. The advantage of manipulability accounts is that they provide a natural way to recover the asymmetry of causality: in a means-end relation, the end obviously depends on the means but not vice versa. The traditional worry with manipulability accounts is that they may be unable to avoid circularity ('manipulating' is a causal notion) and/or anthropocentricity (causality depends on manipulations performed/performable by humans).

4.4.1 Agency accounts

Agency accounts are not well suited to analyse the meaning of causal claims in complex systems. This can be illustrated with reference to Menzies and Price

(1993)'s agency account (in short, AG). Causality is "something analogous to a secondary quality" (Menzies and Price, 1993, p. 189), like colours: to be 'causal' is to trigger some sort of response in an agent in standard conditions, just like to be 'red' is to have the disposition to look red to a normal observer under standard conditions. The analysis reads as follows (see Menzies and Price, 1993, p. 187):

> **[AG]** Event A is a cause of a distinct event B iff bringing about A would be an effective means by which a free agent could bring about B

In this way, Menzies and Price hope to reduce the notion of causation to the more familiar notion of free manipulation. Realising a causal relation between A and B means increasing the probability of B by means of bringing about A.[57] The purported advantage of AG is to account more easily for spurious relations: If A raises the probability of B under manipulation, arguably the relation is not spurious; conversely, if the correlation between A and B (e.g., barometer reading and rain) is spurious, arguably manipulating A makes the probability-raising dependence between A and B disappear.[58]

Menzies and Price claim that their account is non-circular because the notion of manipulation is understood with reference to our *direct experience* of acting as agents doing one thing in order to achieve another, without a prior understanding of the notion of cause. Also, they propose to define 'causes' in the presence of causality *by ostension*, in a way analogous to the way we would define 'red' in the presence of a patch of red, by pointing to it. However, it is debatable whether AG succeeds in avoiding circularity and anthropocentricity.[59]

[57] The probabilities in question are called "agent probabilities", i.e., "conditional probabilities, assessed from the agent's perspective under the supposition that antecedent condition is realized *ab initio*, as a free act of the agent concerned" (Menzies and Price, 1993, p. 190). Notice that Menzies and Price disagree on the interpretation of the agent probabilities: Price holds that the probabilities in question are evidential; Menzies, instead, believes they represent objective conditional chances. As a consequence, arguably they also disagree as to whether their account merely specifies *subjective* assertibility conditions for causal claims (cf. Price, 1998) or also objective assertibility conditions (read: *truth* conditions), like other difference-making accounts. For more on Price's position in relation to the inferentialist account, see §7.2.2.

[58] For a criticism of this claim, see Woodward (2008, §4).

[59] Woodward (2003, p. 125), for instance, argues that Menzies and Price avoid non-anthropocentricity only at the price of circularity.

In particular, even assuming that the notion of agency can be non-circularly used to analyse the notion of causing, AG may not be very illuminating as regards the meaning of causal claims in complex systems, and more generally the meaning of 'causes' in scientific contexts. The problem, is that *analysing* causal relations and *experiencing* or *denoting* them are two different things, in the same way that analysing the meaning of 'red' and experiencing or denoting red are different. So, our 'first-person' understanding of 'red' might well be linked to our own experience of redness but does not say much about the relation between the wavelength of the light emitted by the objects and our sensory apparatus. The latter may be more informative than the former with regard to the meaning of 'red'. Analogously, the phenomenal experience of bringing about is not enough to analyse the meaning of 'causes'. First-person experiences, such as experiencing pressure, are not very informative as regards the meaning of 'causes' in, e.g., 'p53 causes Casp3' or 'chartist behaviour causes volatility'. In these cases, talk of p53 'activating' Casp3 or chartist behaviour 'destabilising' market prices seems more informative.

As a result, AG fails to reduce 'causes' to non-causal notions. We need criteria which are more informative and more faithful to scientific practice than first-person experiences on the basis of which causal claims can be inferred.

4.4.2 Interventionist account

To overcome the above objections, the interventionist account was developed. This account takes inspiration from the literature on causal discovery and inference (Spirtes et al., 1993; Pearl, 2000) and has been developed into a philosophical account of the meaning of causal claims by Hitchcock (2001) and Woodward (2003). I will, from here onwards, mainly refer to Woodward's version of the interventionist account, in short INT, where the notion of causality is characterised in terms of 'ideal' interventions.

A causal relation between variables X_i and X_j is characterised with respect to a causal model, defined by the ordered couple $< V, E >$, where V is a set of causal factors, $V = \{X_1, X_2, ..., X_n\}$, and E stands for a set of n structural equations describing the structure in which the variables are embedded. Usually one distinguishes between 'endogenous' factors X, whose value is set by mechanisms in the structure, and 'exogenous' factors U, which stand for outer influences (e.g., interventions) and/or error terms, and whose value is not set by mechanisms in the structure. Then, the model is defined by the

triple $< U, V, E >$, where E indicates the value of endogenous variables as a function of variables in both U and V. A causal structure is so represented:[60]

$$X_1 = U_1 \tag{4.1}$$
$$X_2 = f_2(X_1) + U_2 \tag{4.2}$$
$$X_3 = f_3(X_1, X_2) + U_3 \tag{4.3}$$
$$... \tag{4.4}$$
$$X_n = f_n(X_1, X_2, ..., X_{n-1}) + U_n \tag{4.5}$$

Causality between a r.h.s. variable and a l.h.s. variable is analysed in terms of 'interventionist counterfactuals', whose antecedent stands for the event where the value of the r.h.s. variable is set by an 'ideal' intervention:

[INT] X_i causes X_j iff an ideal intervention on X_i would make a difference to X_j.

A relation is causal iff an ideal intervention on the cause would make a difference to the effect. 'Ideal interventions' are defined as follows. Let X_i, X_j and I indicate variables, the former two being endogenous, and the latter exogenous. An ideal intervention I on X_i with respect to X_j is such that (cf. Woodward, 2008, §6):

[INT$_1$] I must be the only cause of X_i;

[INT$_2$] I must not cause X_j via a route that does not go through X_i;

[INT$_3$] I should not itself be caused by any cause that affects X_j via a route that does not go through X_i;

[INT$_4$] I leaves the values taken by any causes of X_j—except those that are on the directed path from I to X_i to X_j—unchanged.

[60] Notice the analogy of this kind of representation with the BN-style representation of causal structures as systems of equations of the form $X_i = f(PA_i, U_i)$ (see §4.3.2).

Conditions INT_1–INT_4 ensure that any change in X_j following to the intervention is to be ascribed to X_i: I disrupts the causal relationship between X_i and its previous causes, so that the value of X_i is set entirely by I; I has no direct effect on X_j; I is not caused by any of X_j's causes that are not on the route tested; I does not affect causes of X_j that are not on the route tested.

INT purports to deal with both circularity and anthropocentricity. First, notice that INT does not provide a conceptual reduction of causality, since interventions are defined in causal terms. Yet, INT provides a non-circular, informative characterisation of any target causal relation in non-causal terms—causal terms are only used with reference to *other* causal relations. With respect to the above model, the relation between variables X_i and X_j taking certain actual values $X_i = c_@$ and $X_j = e_@$ is causal iff it is 'invariant' under some interventions, i.e., $e_@ = f(c_@)$ and there is an intervention $X_i = c_{\text{int}}$ which fulfils conditions INT_1–INT_4 and changes the value of X_j to some $e_{\text{int}} = f(c_{\text{int}})$.[61] So, as was for CC, at the level of specific causal claims INT does provide a sort of conceptual reduction. Secondly, although INT mentions interventions, the account is non-anthropocentric, since causality is defined in terms of interventions that need not be carried out *by humans*—the values of the variables may be spontaneously modified by Nature itself—ideal interventions only need to be 'in principle' possible for the counterfactual claim to be true.

The interventionist account has the virtue of making clear the connection between causation and one specific kind of intervention. Still, it has limitations when used as an analysis of the meaning of causal claims in complex systems. In general, it is not clear whether INT is to be interpreted as a *conceptual analysis*—telling how 'causes' is or ought to be used (Woodward, 2003, pp. 7, 132)—or a *methodological criterion*—telling how one is to find out whether causal relations obtain (Woodward, 2003, pp. 8, 22, 114). Either way, the account has weaknesses: if the criterion is methodological, it doesn't apply to non-ideal settings, so it is not an all-encompassing regulative principle; if the criterion is conceptual, it does not capture the meaning of causal

[61] Analogously, in the BN terminology, one may think of an intervention that sets the value of X_i to some $X_i = c_{\text{int}}$, fulfils conditions INT_1–INT_4, and changes the probability of X_j to some $P(X_j|X_i = c_{\text{int}}) \neq P(X_j)$. As mentioned in §4.3.2, the advocates of the BNs approach appeal to interventions to deal with violations of faithfulness. The basic idea is that, although the probabilistic difference does not show up in the original graph, it does in the graph where one sets the value of the variables one by one by intervention and then observes what happens (Pearl, 2000) or considers an augmented graph where intervention nodes on all the variables of the original graph are added (Twardy and Korb, 2004; Korb and Nyberg, 2006). In this sense, INT counts as a modified probabilistic account of causation.

claims in cases of unmanipulable causes, so it does not provide an exhaustive analysis.

First horn first. Woodward argues against Lewisian versions of the counterfactual account: judgements of similarity based on Lewis' possible-world semantics are both unintuitive and imprecise, and this makes causal ascription often problematic (Woodward, 2003, pp. 137-139). In contrast, the interventionist account purports to be intuitive and precise: counterfactuals are evaluated based on judgements about actual or hypothetical experiments, in line with scientific practice. However, under several respects his account has no clear or strong link with methodology.

To begin with, INT offers no all-encompassing criterion for *testing* causal relations. For instance, it is often not satisfied in complex systems. These are characterised by a lot of interactions—or 'couplings', using the complex systems scientists' jargon—both amongst their parts and between the parts and the environment. Hence, there may not be interventions that modify X_i without also modifying X_j either directly or indirectly.

Consider the case of apoptosis. Apaf1 promotes apoptosis by activating Casp9, which then activates Casp3. Above a threshold, stimulation of Apaf1 makes the Casp3 activation irreversible. Below the threshold, instead, Apa1-associated active Casp9 is inhibited by XIAP and unable to trigger Casp3. So, XIAP typically *inhibits* Casp3 activation. However, XIAP also *promotes* Casp3 activation, by contributing to irreversibility. This is because XIAP has a dual capacity: it can also bind to Casp3 (figure 3.3). When Casp3 binds to (non-mutant) XIAP, XIAP cannot bind to, and inhibit, Apaf1-associated Casp9, which is then free to trigger Casp3. Therefore, not only does XIAP decrease Casp3 by inhibiting Casp9; it also increases Casp3, by inducing the implicit positive feedback of Casp3. However, there is no intervention on XIAP by which we could modify Casp3 along the latter path without also affecting variables (Casp9) along the former path, hence without violating INT_2 and INT_4. Notice that, depending on the context (e.g., Casp3>Casp9>XIAP), the second activity may be negligible with respect to the first. However, even if we fix the context so that the *net* effect on Casp3 is either negative or positive, there still are—strictly speaking—*two* activities in operation (figure 3.2, bottom left). Also notice that Legewie et al. (2006) do consider the effect of an intervention that mutates XIAP so that it non-competitively binds to both Casp3 and Casp9. As a result, (mutant) XIAP affects Casp3 without interfering with Casp9. But this intervention, too, violates INT_2 and INT_4. In

fact, it produces *two* different species, namely XIAP-BIR2, with the capacity to bind to Casp3 only, *as well as* XIAP-BIR3, with the capacity to bind to Casp9 only. So, the intervention has effects on *both* Casp3 and Casp9.

Nor is INT very useful when it comes to *use* causal knowledge which one has acquired by means of ideal interventions. Cartwright and Efstathiou (2007), for instance, agree with Woodward that invariance is sufficient to establish causation. If one can perform an ideal experiment where, thanks to background causal knowledge, one is able to control for confounders and test a single dependence relation, one can tell reliably whether the relation is causal. Yet, knowledge of invariant relations so acquired can hardly be *used*. Nothing guarantees, in fact, that changes in the causal structure of the system have not occurred from the time of the experiment. This Cartwright and Efstathiou dub the problem of 'unstable enablers'.[62] Consider the asset pricing mechanism. Let us assume there is a way to intervene on frequency of updating of trading strategies, or on chartist behaviour, to affect volatility. Or let us assume that there is a handle, e.g., 'network connectivity' (Anand et al., 2011), on which it is possible to intervene to prevent crashes in the light of 'precursory fingerprints' (Sornette, 2002). If we are lucky, these interventions may be successful. Still, due to reflexivity, they would also modify the system, so that the exploited functional relation would not remain invariant. In Cartwright's words, "our actions can undermine the very structure that gives rise to the causal principles we rely on to predict the outcomes of our actions" (Cartwright, 2007b, p. 40). Economists themselves admit that social policies inevitably trigger changes in structure that demand revision of our knowledge of the structure and possible policies. They regard 'ideal' policies not as interventions that leave the relations invariant, but as interventions that manage to achieve short-term goal, in spite of the evolving nature of the system (see Kirman, 2010, pp. 526-527). Accordingly, they do not search for optimal solutions, viz. solutions of the kind ideal interventions allow, but for 'satisficing' solutions, and then adapt their policies on-the-fly to changes in structures (cf. Simon, 1996, chap. 2). In the light of this, INT is of scarce utility for policy, hence *not* methodologically necessary.

But perhaps the criterion in INT should not be so strictly interpreted.

[62] Assuming also modularity (i.e., that each direct relationship among variables can be intervened upon without disturbing the others), as Woodward does, guarantees good predictions in the system tested, that is, *internal* validity. However, assuming modularity may be bad for scope: modularity may not hold across systems, or even for the same system at different times—which is a problem of *external* validity.

So far, it has been assumed that the right causal structure is known, with the exception of the causal relation to be tested. In fact, knowledge of the right causal structure is required to perform surgical experiments and control for confounders. In practice, however, one doesn't have this knowledge. Given this difficulty, one often has to rely on other tools and assumptions, e.g. RCTs. Couldn't one interpret ideal interventions in INT as a (merely) "regulative ideal" (Woodward, 2003, p. 114)? On this reading, what Woodward may be saying is not that ideal interventions are a sort of all-or-nothing criterion for causation, but rather that the more ideal the intervention, the more warranted the causal claim. For instance, INT may be approximated by an average causal effect criterion (§4.3).[63] This criterion, common in the statistics literature (see, e.g., Holland, 1986), says that a cause raises the *overall* probability of the effect, in the sense that if the putative cause is appropriately manipulated positive differences in some populations outrun negative differences in other populations. RCTs could be interpreted as attempts to *approximate* ideal interventions. Take two populations of traders which, by randomisation, are alike in all respects but the value of some control parameter X_i, which is set by a (non-surgical) intervention in the test population and left unchanged in the control population; even in the absence of knowledge of the exact causal structure, provided the confounding effect of B on X_j has been eliminated by randomisation, any probabilistic difference on X_j can be ascribed to the causal effect of X_i.

The problem with this interpretation of INT is that for an RCT to establish causation, INT_1–INT_4 need not be satisfied. Stability under (non-ideal) interventions, viz. robustness across changes in context, is enough to establish causation. For Woodward, stability matters (only) to the *scope* of the causal claim, that is, to whether a causal claim counts as a causal 'law': 'X_i causes X_j' is a causal law if the counterfactual dependence of X_j on X_i holds across a range of values B of other variables in the background, $X_j = f(X_i, B)$; invariance, he maintains, is necessary for the relation to count as causal. But in a well-conducted RCT, it is overall probability raising that matters; invariance over each unit is methodologically unnecessary. —For instance, a treatment that violates INT_4 but makes a *net* difference to the effect, still

[63] The *causal effect* on X_j of a change in X_i from $X_i = x_{@}$ to $X_i = x_{\text{int}}$ (against a stable background context of values of the other variables) is defined as the counterfactual difference between the value that X_j would take under the intervention $X_i = x_{\text{int}}$ (viz. the *counterfactual* value of X_i) and the value that X_j would take under the intervention $X_i = x_{@}$ (viz. the *actual* value of X_i).

counts as a cause.[64] So, Woodward's regulative ideal is not a good guide in *all* circumstances.

If the above objections are sound, there is *more* to causality than envisaged by INT. These and other limitations have led some (e.g., Cartwright, 2007b) to complain that the interventionist account is too often unable to represent causal relations, so cannot constitute a complete analysis of causality. The interventionist account is, like the other accounts reviewed in chapters 4 and 5, 'mono-criterial': it privileges one criterion for causation over other criteria.[65] Woodward thinks that this mono-criterial character is a problem for INT *only if* there are realistic cases in which INT and other criteria conflict and where it is clear that the causal judgements supported by these other criteria are more defensible than those supported by INT (Woodward, 2008, §14). However, contrary to Woodward, I believe that the tenability of INT does *not* depend on the existence of cases of causation such that some other monistic analysis is more suitable than INT to account for them. It suffices that there are cases, such as the aforementioned apoptosis and asset pricing examples, where our intuitions converge on the judgement that there is causation and that one or more of the assumptions in INT fails—irrespective of the availability of an alternative analysis.

Let us now address the other horn: to what extent does INT provide a *conceptual* analysis of causality? A widely-held view on meaning is that the meaning of a concept depends on its contribution to the truth conditions of the sentences in which the concept appears. So, on the implicit assumption that whatever provides truth conditions constitutes a better conceptual analysis than what doesn't, Hiddleston (2005) and Psillos (2004) argue that laws provide a better analysis of causal claims than INT: whereas laws provide truth conditions[66], interventionist counterfactuals provide at most test conditions. In fact, what interventions count as ideal depends on the invariance of the relation; invariant relations, in turn, are defined in terms of counterfactuals; thus, so they argue, *if* counterfactuals are ultimately reducible to laws, the meaning of causal claims can be analysed in terms of laws only, without

[64] Cartwright offers a similar remark to the point that neither invariance nor other test conditions are necessary to use causal claims for the task of prediction. What matters is the relation's *stability* across contexts (see Cartwright, 2007b, p. 50).

[65] In contrast, Cartwright (2007b) favours a *pluralist* account over monistic, or mono-criterial, accounts. A pluralist believes that a variety of distinct criteria may be relevant depending on the causal claim at issue (see chapter 6).

[66] Notice that laws are not the only candidate. An alternative is to take mechanisms as truth-makers of causal relations (see chapter 5). For more options, see Strevens (2007).

reference to interventions. Woodward's reply is that this objection relies on the unjustified assumption that one can give a reductive account of laws (or mechanisms) in non-counterfactual terms (Woodward, 2008, §14).

I agree with Woodward on this. However, the moral I'd like to draw is more general, and extends to the concept of causation as well: we should not try to reduce 'causes' to other concepts, only limit ourselves to point to its connections with other concepts. That INT cannot function as a conceptual analysis of causation is shown by illustrating its limitations in dealing with cases where we have no clear intuition on the objective conditions for the outcomes of hypothetical experiments. Large attention has been drawn to outcomes which depend on logical/conceptual possibilities that involve *violation* of laws. Take the claim 'The position of the moon influences the tides', discussed by Woodward (2003, pp. 130-131). Changing the position of the moon by doubling, for instance, its orbit without affecting the tides, in a way or another, would require a violation of physical laws. And what is the outcome of such experiments—and the truth value of the corresponding counterfactual claims—in a world where not only initial conditions but also laws are changed? However, no sufficient attention has been drawn to the fact that violations of laws are far from being the only case where we lack clear intuitions. In many cases (e.g., apoptosis and asset pricing), the non-intuitive character of the outcomes of the experiments depends not on violations of laws but rather on the complex and sensitive nature of the relations among the causal factors.

An alternative is to read INT as providing an analysis of our causal *intuitions* not of causal *relations* themselves. Woodward claims that the function of INT and of thought experiments in cases where the intervention is only logically-conceptually possible is to "give us a purchase on what we mean or are trying to establish when we claim that X causes Y" (Woodward, 2003, p. 130). So, it seems plausible to interpret him as giving a criterion for *meaningfulness* not truth conditions: the relation between two variables is *conceivable as* causal iff we can devise a thought experiment in the form of an ideal intervention on one variable by which to modify the other variable. This interpretation is considered by Psillos (2004, p. 301), too, who then discards it for the following reason: although it is plausible to take INT as providing sufficient conditions for meaningfulness of causal claims, it is doubtful that it also provides necessary conditions. Take the moon-tides case again. The meaningfulness of the causal claim does not seem to rely on the existence of

clear intuitions as regards what counts as a conceptually ideal intervention. In fact, it is unclear which interventions are conceptually legitimate and which aren't. As a result, it is unclear how INT can decide which causal claims are meaningful and why.

In general, for Woodward INT is informative on the meaning of causal claims as long as it provides a "principled basis" to answer counterfactual questions about what would happen to the value of some variable if an intervention were to occur on another variable (Woodward, 2008, §6). However, it is not clear whether the counterfactuals that are relevant to the meaning of causal claims must *always* be understood in terms of interventions—whether explicitly or implicitly. As Woodward himself admits,

> as we make the relevant notion of "possible intervention" more and more permissive, so that it includes various sorts of contranomic possibilities, we will reach a point at which this notion and the counterfactuals in which it figures become so unclear that we can no longer use them to illuminate or provide any independent purchase on causal claims (Woodward, 2008, §11).

An analogous reasoning applies to a causal claim involving two successive states of the entire universe where one is supposed to cause the other. Here, like in the case of two successive states of a system which is completely isolated from outer influences, interventions are in principle impossible. However, interventions in the universe case are impossible for a more fundamental reason, namely not only because the system is totally shielded from what is outside, but also because there is nothing outside the causative state which could be used to realise an intervention on such a state. Woodward (2008, §12) reads this as an indication that INT applies more naturally to small worlds, where exogenous interventions are possible. Additionally, one may draw the following conclusion (C): The larger the world considered, the less the meaning of a causal claim *has to do with interventions*. Instead, Woodward concludes (C'): The larger the world, the less a causal claim *is meaningful*. But I see no reason why we should prefer C' over C. On the contrary, assume one is presented with two causal claims, one about two successive states of the universe and one about two successive states of an isolated system. Intuitively, it makes as much sense to say that a state of the universe causes a later state of the universe as it makes sense to say that a state of an isolated system causes a later state of the system. Furthermore, the legitimacy of the causal

intuition in the former case seems independent from considerations involving the possibility, whether practical or conceptual, of interventions.

In general, the more often INT is not useful or illuminating, the less plausible it is that INT provides an *exhaustive* account of causation. One is then left to wonder what makes causal claims *causal*, i.e. what is their meaning, when INT is not clearly applicable. In such cases, it seems more plausible that causal claims get their meaning from the analogy they bear with other claims, from considerations of simplicity and coherence, etc. rather than from the existence of ideal interventions that meet INT_1–INT_4, or from judging on how much real interventions approximate ideal ones.

4.5 The contextuality of causality

Discussion of difference-making accounts supports the conclusion that the meaning of 'causes' is not reducible to necessary and sufficient conditions. Although each account has virtues, and is clearly informative in some domains, it does not apply so well to other domains. For instance, RT is typically informative where causal relations happen to be embedded in complex nets of linear relations; not so informative when effects are sensitive to initial conditions or determined by many nonlinear interactions. CD is typically informative in the presence of clear intuitions about possible worlds, and in the absence of chaos, overdetermination or preemption; not so informative otherwise. BNs are typically informative when probabilistic dependences are either intuitively causal or some set of causally relevant variables happen to screen them off, and when causal relations in turn faithfully generate probabilistic dependences and independences; not so informative when there are loops, and PCC or causal faithfulness are violated. INT is typically informative when INT_1–INT_4 are satisfied; not so informative where relations are intuitively causal even though one or more of INT_1–INT_4 are violated.

In sum, the conditions from which a causal claim can be legitimately inferred, or *test* coditions, and the consequences that a causal claim entitles one to infer, or *use* conditions, may vary from context to context. Accordingly, the notions which best serve to explicate the meaning of 'causes' vary as well. The problem with complex systems is that, in addition to the general issue of contextuality *as domain sensitivity*, here the meaning of 'causes' is not only domain-relative but also *claim*-relative. Even if we focus on this *one* specific domain (in fact, on only two mechanisms), no analysis seems well-suited to

describe causal relations in it. What moral should be draw from this?

In the following, I argue that one should not take contextuality as entailing that no informative analysis of the meaning of 'causes' is possible, or that causal claims are irreducibly subjective. The detailed argument is given in chapters 6 through 8. Here is a sketch of my argumentative strategy.

First, the context-sensitivity of the meaning of 'causes' is not incompatible with the possibility that all instances of 'causes' be instances of a common concept (see chapter 6). However, this possibility need not rely on the reducibility of the meaning of 'causes' to necessary and sufficient truth conditions. A claim to the contrary, on the mere ground that an anti-reductionist view about causality is uninformative, is question-begging (cf. Carroll, 2009, p. 290). Nor is the truth-conditionalist analysis of meaning the only game in town. Here is an alternative, which I investigate in chapter 7: the meaning of causal claims, and the word "causes" in such claims, comes not from their correspondence with mind-independent facts, whether causal or non-causal, but from their *functional role* in the language game, that is, on their being (parts of) premisses or conclusions of arguments (Harman, 1999; Dummett, 1991; Brandom, 1994b).

Secondly, the contextuality of causal claims is not incompatible with their objectivity, provided 'objectivity' is not identified with 'mind-independence' of truth conditions, but with 'entitlement' to the claim on the basis of practice, both linguistic and non-linguistic. Objectivity, so construed, is partly a matter of linguistic rules and norms, and is not reducible to some alleged realm of facts describable in mind-independent terms. I leave to chapter 8 the task to justify the view that, although in a sense causal claims are sensitive to the context of their use, in another sense they have a force which goes beyond such a context.

To pursue this argumentative strategy, I will resort to an inferentialist analysis of the meaning of causal claims. As I will argue, inferentialism allows one to make sense not just of the contextuality of causal claims (their meaning varies from context to context) but also of their objectivity (relative to the context, one claim is more or less appropriate).

Conclusion

Difference-making accounts provide useful test conditions for causality. Yet, they do not constitute exhaustive analyses of the meaning of causal claims

in complex systems. On the one hand, difference-making criteria are not always satisfied, due to sensitivity to context and initial conditions. On the other hand, due to their almost exclusive focus on test conditions, difference-making accounts disregard the role of *use* conditions in establishing meaning. The first take-home message of the chapter is that the meaning of 'causes' has different connotations in different contexts. A satisfactory account of causality in complex systems should both *acknowledge* and *explain* such a contextual aspect: on the one hand, it should say what 'causes' means in cases where specific difference-making criteria do *not* apply; on the other hand, it should explain the relevance of difference-making criteria where/when they *do* apply. The second take-home message of the chapter is that analyses of the meaning of causal claims in terms of necessary and sufficient truth conditions may be wrongly-headed. It is time to try something different. Before turning to my positive proposal, however, I need to show that other accounts of causality are not well suited either to explicate the meaning of causal claims in complex systems, namely mechanistic accounts (chapter 5) and pluralist accounts (chapter 6).

Chapter 5

Mechanistic Accounts of Causality

In this chapter, I discuss the so-called 'production' accounts of causality. These accounts are based on the intuition that a cause is something that produces, or brings about, the effect. The explication of the notion of causality in such accounts goes via the identification of the truth-maker of the relation, whether a process, a disposition, a mechanism, or else. My main focus here will be the mechanistic account developed by Glennan (1996, 2002), since this is explicitly meant to fit the complex systems case. In §5.1, I introduce the reader to the motivations for the mechanistic approach in philosophy of science and in the analysis of causality. In §5.2 and §5.3 I present, respectively, Glennan's account and some objections against it. In §5.4 and §5.5, I discuss the prospects of the mechanistic account. In particular, I argue that the account cannot be salvaged by drawing on conceptual resources from other production accounts, viz. process-based accounts and power-based accounts. I conclude with a remark on the possibility to give a non-representationalist interpretation to models of mechanisms and to the causal claims describing the mechanisms' workings (§5.6).

5.1 Mechanistic causality: whys and wherefores

'Complexity' and 'mechanism' are deeply intertwined notions: understanding complexity requires an understanding of the mechanisms which produce and sustain it, and many mechanisms are—at least *prima facie*—complex. It is no surprise, then, that talk of mechanisms is ubiquitous in complex systems sciences. Upon realising that a given model applies well to a given class of systems, scientists tend to say that the systems belonging to the class instantiate the same—or, at any rate, a similar—*mechanism*. Also, complex systems sciences are often cross-disciplinary, since features common to a class of, say, biological systems may be shared by certain social systems as well. Indeed,

what scientists often realise is that, surprisingly, abstract mathematical or computational models which apply well to a class of systems are 'exportable' to *prima facie* very different systems. For instance, the applicability of network models (Newman et al., 2006) ranges from protein-protein interactions to the world wide web; models of self-organisation (Nicolis and Prigogine, 1989; Kauffman, 1993) can be applied to phenomena as diverse as convection and magnetisation on the one hand, and market behaviour on the other hand; genetic algorithms (Holland, 1995) can be used to model biological evolution as well as the evolution of decision-making strategies, e.g., in minority games; and so on and so forth. In such cases, scientists (and philosophers, too) find themselves asking what *mechanism*, if at all, is shared by two different classes of systems which makes the transfer from one class to the other successful.

Since talk of causal relations in complex systems sciences is usually associated with descriptions of mechanisms, a plausible working hypothesis is that causal relations in complex systems have somehow to do with mechanisms and that a satisfactory account of causality in complex systems should be informative as to the connection between mechanisms and causal relations. And clearly, even if existing mechanistic accounts of causality are *not* satisfactory for one or the other reason, it is desirable to understand *why* it is so, and build a better account from there.

At the same time, philosophers interested in scientific methodology, in particular scientific explanation, are more and more focussing on understanding the nature of mechanisms. Besides, some philosophers have come to conceive causal relations themselves as 'requiring' mechanisms.[67] In general, philosophers wishing to offer a mechanistic account of causality have to consider at some point the role that complexity plays in shaping causal relations. This is all the more true if the mechanistic account of causality is explicitly meant to fit complex systems causation, as is the case of Glennan (1996, 2002). The reason for this interest in mechanisms in the causality literature is that mechanisms are considered crucial to satisfy the following two *desiderata*: (i) to provide the (causal) *explanation* of phenomena; and (ii) to elucidate the concept of causality by identifying the *truth-makers* of causal claims.

Mechanistic accounts of causality purport to meet the first, explanatory requirement by showing how mechanisms provide the thread between putative causes and effects. The idea is that the effect happened because certain

[67] Here, 'requiring' may be read in several ways, viz. as 'being instantiated in', or 'being analysable in terms of', or 'being explainable in terms of', or what have you.

conditions and a certain mechanism were in place. The explanation is usually supposed to avoid appeal to magic or action-at-a-distance. It is thought that magic or action-at-a-distance don't provide the desired explanation, whereas mechanisms 'open the black box'.

The second requirement, instead, concerns the informativeness of the account. Even if the aim of the mechanist is not to give a conceptually reductive analysis of causation, he still wants to produce an informative characterisation of causality, by saying what all causal relations have in common—although this characterisation may be in terms of notions with *causal* content (e.g., 'produces', brings about', 'inhibits', etc.).

The mainstream view on what a mechanism is rests on the definitions in (Machamer et al., 2000; Bechtel and Abrahamsen, 2005; Glennan, 2002). Despite differences, all parties agree that mechanisms consist of entities/parts, their activities/interactions, and their organisation, which are together responsible for the production of the phenomenon. For the sake of precision, other characterisations should be added to the above list.[68] Most characterisations, however, either constitute mere variations on the same theme, or are less relevant to the case of complex systems causation. Thus, with the exception of the interventionist account of mechanisms (§5.3), the Salmon-Dowe account (§5.4.4), and Cartwright's 'nomological machine' account (§5.5.1), they will not be discussed in the present context. Indeed, I will mainly focus on Glennan's account, since it is the only mechanistic account which explicitly purports to be an account of *causality* and to apply to *complex systems*.

5.2 Glennan's mechanistic account of causality

Glennan (1997, 2011) adopts a 'singularist' view (cf. §4.2): the obtaining of the causal relation is *local* to the occurrence of cause and effect and what is between them. Causal *laws* are to be understood as descriptions of, and derivative from, singular facts about the local workings of mechanisms. This

[68] For instance, Salmon (1997, pp. 462, 468) and Dowe (1995, p. 323) characterise mechanisms in terms of processes and interactions where physical quantities are conserved (§5.4.4). Woodward (2002, p. S375), in line with his interventionist account, defines a mechanism as a structured set of parts such that its behaviour is describable by means of generalisations which are invariant under ideal interventions. For Bunge (2004, pp. 189, 191, 193), mechanisms are processes in material systems; processes, in turn, are sequences of states, or strings of events. Cartwright (1999, p. 50) holds that mechanisms are a sort of 'nomological machines', that is, arrangements that tend to produce the regularities described by scientific laws. Elster (1989, 1998), Hedstrøm and Swedberg (1998) and Steel (2004) offer other characterisations which are meant to be tailored to *social* systems.

view is incompatible with analyses of causation based on generalist criteria, e.g., regularity and probabilistic dependence accounts.

According to the mechanist (see, e.g., Glennan (1996, p. 64), Bunge (2004)), a relation between two events is causal only if it is underwritten by some mechanism which connects them. It is well known, however, that there are all sorts of mechanisms between any two event types. For instance, between sunlight and cancer there are mechanisms that bring about cancer (excessive sun exposure) as well as mechanisms that prevent it (given that, trivially, the whole of life depends more or less directly on sunlight). The problem, then, is how to pick the appropriate mechanism, that is, the mechanism that explains and provides the truth-maker for the relation. Glennan's first attempt to achieve these goals by spelling out the notion of mechanism resulted in the following definition:

> A mechanism underlying a behavior is a complex system which produces that behaviour by the interaction of a number of parts according to direct causal laws (Glennan, 1996, p. 52).

However, given the absence of regularities unanimously accepted as laws in complex systems sciences, he later changed this definition so as to avoid reference to the universality commonly associated with laws:

> A mechanism for a behavior is a complex system that produces that behavior by the interaction of a number of parts, where the interactions between parts can be characterized by direct, invariant, change-relating generalizations (Glennan, 2002, p. S344).

The general idea behind the above definitions is that not all systems will count as mechanisms, but only those which are stable (e.g., watches, cells, organisms, social groups), made of parts which are themselves stable, and whose interactions produce a robust behaviour, as opposed to an ephemeral one (see Glennan, 2002, p. S345).

The change from the former to the latter definition constitutes a shift from analysing causation in terms of laws, that is, counterfactual-supporting generalisations (see Glennan, 1996, p. 54), to regarding such generalisations only as a way to describe the relation between facts about (ideal) interventions and facts about outcomes of these interventions.[69] The requirement that they

[69]This shift was anticipated by (Glennan, 1997), where any (non-fundamental) law,

be 'direct' is meant to avoid intervening causal factors along the pathway from cause to effect.

Following Woodward, Glennan says a relation is causal when there is a possible intervention that by modifying the value of a variable (which corresponds to the property of one part) brings about a change in the value of another variable (which corresponds to the property of another part) without altering the values of the other variables (see Glennan (2002, pp. S344-S345), Glennan (2011, §4) and Woodward (2003, p. 98)). Such an intervention helps identify generalisations which are 'change-relating', i.e. functionally relating changes in the cause and corresponding changes in the effect, and 'invariant', i.e. stable across a range of values of the other variables. This, in turn, helps distinguish causal relations from accidental correlations or spurious relations between effects of a common cause.

Glennan's account differs from Woodward's, however, because causation is not *analysed* in terms of interventions. Interventions provide test conditions rather than truth conditions. What makes the relation causal is the mechanism linking the cause and the effect, not facts which merely *exploit* the existence of this mechanism. In Glennan's own words, "The manipulability account emphasizes procedures for discovery, prediction and control. The mechanical account provides (...) a metaphysical underpinning of the manipulability approach" (Glennan, 2011, p. 802).

The account is *prima facie* attractive since, as I said, it purports to provide a characterisation of causal relations *in complex systems*. Also, it appears to have the resources to achieve this goal, as it is cashed out in terms of notions which complex systems scientists themselves use, namely, parts, interactions, 'complex' systems, etc. As we shall see, however, it suffers from problems and ambiguities which arise when the notion of mechanism is used to characterise causal relations.

5.3 Problems with the mechanistic account

Glennan's characterisation of mechanism goes via the notion of counterfactuals. Against his proposal the objection could be levelled that it gives rise to circularity and/or regress, both of which may result in damaging the informativeness of the account.

e.g. Mendel's second law, is characterised as "just a description of how a particular type of mechanism behaves" (*ibid.*, p. 622).

Mechanisms that underpin causal relations are characterised in terms of a particular kind of counterfactuals, namely *interventionist* counterfactuals, which are used to produce change-relating generalisations. Each generalisation, in turn, describes the operation of another mechanism. So, the account seems *conceptually circular*—causality being analysed in terms of mechanisms; mechanisms in terms of counterfactuals; counterfactuals, in turn, in terms of mechanisms; and so on and so forth. Notice that this need not be a problem in itself. As was the case of non-reductive analyses of 'causality' in terms of probabilities and ideal interventions, here, too, the analysis may be informative, provided the circular connection between 'mechanism' and 'intervention' is virtuous. (One obvious difference is that, whereas in the probabilistic and interventionist analyses the analysans provides a *direct* analysis of 'causality', in the mechanistic account it provides an *indirect* analysis, since the link to 'intervention' goes via the notion of 'mechanism'.) Circularity *is* a problem, however, if the circle fails to shed light on the analysandum. Probabilistic and interventionist analyses fail to be informative in cases where the conditions they postulate are not met. The analysis of 'mechanism' in terms of 'intervention' fails to be informative when it presupposes knowledge that the intervention is testing a relation which—actually, or plausibly—belongs to the mechanism. The more one lacks clear intuitions on what the relevant mechanism is, the more 'causality' inherits the vagueness of 'mechanism'. This is often the case in complex systems, where there are no neat/known boundaries between system and environment, and so many operations take place at the same time whose relevance, if any, to the mechanism is unknown.

In addition, between mechanisms and counterfactuals there seems to be an *asymmetry* (see Psillos, 2004, p. 310). Each mechanism is recursively decomposable into parts, each part being a further mechanism, until the system cannot be decomposed anymore into parts. Mechanisms ultimately bottom out in 'brute' counterfactuals, which are therefore more fundamental than mechanisms: (i) fundamental level interactions can only be explained counterfactually ('if this part were to change, that part would change'), not mechanistically, so the above circle bottoms out in brute counterfactuals, which are the ultimate explainers; (ii) the truth-makers of causal claims, whether themselves causal (e.g., Salmon-Dowe processes and interactions) or not, are ultimately *not* mechanisms. Here, the worry is that the account leads to a *metaphysical regress* as regards the identification of the truth-makers: a mechanism at level n is identical to one or more mechanisms at level $n-1$, which

are in turn identical to one or more mechanisms at level $n-2$, and so on and so forth until some fundamental level is reached, *if* such a level exists.[70] Notice that Glennan (1996) takes pain to stress that each mechanistic decomposition has a certain autonomy. That is, the further mechanisms a mechanism can be decomposed into are mechanisms *for other behaviours*, hence they do not provide an explanation for the original phenomenon that we wanted to explain. This may salvage the explanatory function of mechanisms, but doesn't address the issue of truth-makers. In fact, case by case, it would still be true that coarse-grain phenomena, although explainable in terms of coarse-grain mechanisms, obtain in virtue of fine-grain mechanisms—and ultimately in virtue of brute counterfactuals.

And there are other problems, too, concerning Glennan's notion of mechanism. On the one hand, the identification of 'mechanism' with 'system' leads to *unintuitive consequences*. According to Glennan, the *relata* of the causal relation are *events*. Events, in turn, are (causally) related by a mechanism, which is a complex system. A complex system, in turn, is a stable arrangement of parts, which is arguably an *object*. Hence, events would be related by an object. But this sounds implausible. Intuitively, events are mediated by something dynamic, e.g. a process of change, not something static, e.g. an object. How can a system, i.e. an object, provide the thread between two events? As it is, Glennan's notion of mechanism doesn't seem the sort of thing in terms of which causation can be analysed. On the other hand, the characterisation of mechanism seems *too restrictive*. Two objects gravitationally attracting each other, for Glennan, do not count as a mechanism in the sense of a complex system: their interaction is "brute" (Glennan, 1996, p. 50). Between events involving such interactions, the causal relation is characterised in terms of brute counterfactual dependence not a mechanism, since no further decomposition is possible. However, it isn't clear why we have no mechanism here. We do have a system, (two) parts, interactions, etc. This limitation is all the more striking if we consider *prima facie* genuine cases of complex systems, e.g., double pendula, which would not count as a complex system in Glennan's sense. A double pendulum is a complex system whose initial conditions can determine a chaotic behaviour just in virtue of the position of the two masses. This seems a legitimate *mechanistic* explanation, although it does not refer to the operations of further parts. Legitimate seems

[70] If one finds problematic the idea of a 'regress' that could stop somewhere rather than go on forever, one may want to re-label this problem the "bottoming-out problem", viz. the truth-makers at the bottom level, wherever that level is, are not Glennan's mechanisms.

also claiming that the chaotic behaviour is *caused* by the initial position of the masses. However, a double pendulum would not count as a complex system in Glennan's sense, so his account cannot explicate the causal claim. Thus, Glennan's notion of mechanism leaves out too much.

What Glennan would need is a more solid notion of mechanism, which allows to avoid the above mentioned problems, viz. conceptual circularity, metaphysical regress, the unintuitive appeal to objects to mediate between relata, the inapplicability of the notion to two-part mechanisms. In §5.4, I evaluate the prospects of the mechanistic account as applied to complex systems, in the light of more recent developments of Glennan's view as well as considerations of my own.

5.4 The prospects of the mechanistic account

5.4.1 A virtuous circularity?

Recently, Glennan (2011, §4) has advanced the thesis that counterfactuals and mechanisms really are *on a par*—both conceptually and metaphysically:

> The mechanical approach relies on the counterfactual approach because there is no way to define interactions between parts of mechanisms except by appeal to counterfactual-supporting generalizations. The counterfactual approach relies on the mechanical approach because the truth-conditions for counterfactuals depend upon the structure of mechanisms (Glennan, 2011, p. 806).

Glennan grants to Psillos that the mechanical approach cannot eliminate counterfactuals, hence it cannot provide a reductive analysis of causal claims. Yet, he rejects Psillos' asymmetry claim on the grounds that the truth conditions of interventionist counterfactuals depend—more or less explicitly—on the structure of mechanisms. That is, one has to mention the mechanism in which the counterfactual relation is embedded in order to provide the truth conditions of the causal claim, hence one cannot do away with mechanisms either. However, this reply is not, in my opinion, satisfactory.

Leaving aside the regress problem for the moment, which I address in §5.4.3, the (symmetric) interplay between mechanisms and counterfactuals fails to illuminate the meaning of causal claims in complex systems. Here is why. Woodward defines a 'mechanism' as

> (i) (...) an organized or structured set of parts or components, where (ii) the behavior of each component is described by a generalization that is invariant under interventions, and where (iii) the generalizations governing each component are also independently changeable (Woodward, 2002, p. S375).

Depending on how one reads Woodward, one may ascribe to Glennan a stronger or a weaker view on the conditions for interventionist counterfactuals to provide the semantics of causal claims.

Under the strong intepretation *modularity* must hold for something to count as a mechanism, i.e., it must be possible for a change in a property C of one part to bring about a change in the property E of another part without (1) affecting E directly or indirectly or (2) altering any of the other functional relationships in the mechanism (see Woodward, 2003, p. 329). This interpretation is supported by other statements of Woodward's, e.g.:

> the components of a mechanism should be independent in the sense that it should be possible in principle to intervene to change or interfere with the behavior of one component without necessarily interfering with the behavior of others (Woodward, 2002, p. S374).

It seems that Glennan *needs* modularity for his analysis to work. His requirement that the interactions between the parts be 'direct' (§5.3) is intuitively satisfied by a modular system, but makes little sense if independence is not imposed between the relation tested and other equations.

However, since it is not always clear whether Woodward (2003) does in fact claim that causality requires modularity, one may interpret him as making the weaker claim that *at least* (1) must be met. In line with this, he states that "components of mechanisms should behave in accord with regularities that are invariant under interventions" (Woodward, 2002, p. S374).

So Glennan could be interpreted as making either one of the following two claims: (a) 'mechanism' is analysable in terms of 'intervention' whenever mechanisms are modular; or (b) 'mechanism' is analysable in terms of 'intervention' whenever all causal relations in the mechanism are invariant under intervention. However, (1) is often not met in complex systems (§4.4.2): due to the many couplings among the variables, complex systems are often such that there are no interventions that modify E just in virtue of modifying C,

without also modifying E either directly or indirectly. As a result, the semantics of causal claims in such cases remains unclear, since no other criterion besides intervention is offered to distinguish genuine from spurious relations.

5.4.2 What is a mechanism?

In an attempt to elucidate the relation between (mechanical) system and (mechanical) process, Glennan says:

> "Mechanism" is used to describe two distinct but related sorts of structures. First, mechanisms are systems consisting of a collection of parts that interact with each other in order to produce some behavior. So, for instance, a car's engine is a mechanism containing many parts whose interaction produces the motion of the drive shaft. Second, mechanisms are temporally extended processes in which sequences of activities produce some outcome of the mechanism's operation. For instance, photosynthesis is a mechanism which, by a series of activities involving water, carbon dioxide, and energy from light produces oxygen and sugar. There is a natural relationships between processes and systems, for the operations of systems give rise to processes. Photosynthesis can, for instance, be conceived of as the activity of a system—the chloroplast—whose operation is a mechanical process (Glennan, 2008, p. 376).

But this still leaves unclear what a process is[71] and the exact relation between process and object. In this latter regard, Glennan states, quite epigrammatically, that "the operations of systems give rise to processes". However, in the case of complex systems this is not obviously so. It is true that—in a sense—relatively stable systems give rise to processes. In another sense, however, it is processes that give rise to systems (and their operations). In fact, underneath the system's relative stability, there are parts growing or shrinking, entering or leaving the system, etc.

This is true not only for the cell, where processes of metabolism and protein synthesis take continuously place, thereby changing the parts' number and identity. It is true for the market, too. Here, agents may enter and leave

[71] Notice that, as I argue for in §5.4.3, one cannot appeal to the Salmon-Dowe process theory, arguably the most influential account of causal processes, to characterise complex systems' processes. This ultimately leaves Glennan's notion of process undefined.

the system, that is, start and stop trading. Or they can change their attitude towards a given asset, by becoming fundamentalist (expecting the price to follow the 'fundamental value' of the asset, according to the efficient market hypothesis) or chartist (trying to identify and exploit 'charts', viz. trends and patterns), pessimistic or optimistic (Lux and Marchesi, 1999). Or they can adapt their trading strategy on the basis of the asset's performance, continuously hypothesising and refining their expectational models (Arthur et al., 1997). Or... The macro-features associated with asset pricing fluctuations are the result of such micro-processes. Take, for instance, the fat-tailed distribution of returns (which indicates that extreme events, viz. bubbles and crashes, obtain more frequently than if returns were normally distributed) and the volatility persistence of asset returns at different times (which, against the 'random walk hypothesis', display substantial dependences). These statistical features can be reproduced by, e.g., modifying the agents' perception of trend direction and of other agents' profits (Lux and Marchesi, 1999) or by modifying the rate of change of the agents' process of expectation formation (Arthur et al., 1997).

In both cases—the cell and the market—it seems equally legitimate to say that it is processes that give rise to systems, by continuously maintaining and re-shaping them. This, in turn, would fit better with the idea that the relata of the causal relation are events, and that between events there are processes not objects.

5.4.3 Where are the truth-makers?

The unclear status of mechanisms makes the mechanistic account vulnerable to the regress objection. How should we understand the nature of the truth-makers of causal claims? Recently, Glennan (2010, 2011) has expressed two distinct views on this issue.

Glennan (2010) argues that there are two notions of cause, namely causal *production* and causal *relevance*:[72] there can be production without relevance (e.g., overdetermination); and there can be relevance without production (e.g., omissions). He articulates this view by having the production/relevance dichotomy to track the events/properties distinction and the singular/general distinction. Causation is characterised as a relation between events. Events, in turn, are instances (or occurrences, or exemplifications) of properties. Ac-

[72] The dichotomy between production and relevance is reminiscent of Hall (2004)'s two-concept view (§6.3).

cordingly, a causal claim has the following form:

> Event c causes e (in background conditions B) in virtue of properties P (of c, e, or B) (...) [e.g.:] Bob's coughing (c) caused Carol to wake up (e) in virtue of cough's loudness (P) (Glennan, 2010, p. 364).

Causal production holds between causally related events and is not a counterfactual notion. Causal relevance, instead, holds between causally related properties and is a counterfactual notion. But production and relevance are not in opposition, since they serve different purposes. Causal relations *obtain* thanks to mechanisms and *are explained* by them, but in virtue of different aspects of the mechanisms involved. In fact, although mechanisms provide both truth-makers and explanation for causal relations, they do so by relying on different notions of cause: (i) *whether* c causes e depends on whether there is a causal process from c to e ("To say that one event produced another is to say that in fact the causative event is connected to the effect via a continuous chain of causal processes" (Glennan, 2010, pp. 365-366)); (ii) *why* C-type events cause E-type events (of which c and e are instances), instead, depends on some causally relevant feature P of the mechanism, its parts and organisation, and its background condition. So, it seems that whereas difference making is relevant to causal *explanation*, production (processes) is what contributes the truth-makers—which prompts the question: how should we understand 'process', exactly?

Next to this general thesis on production and processes, Glennan (2011) has another, more specific thesis on the truth-makers of causal claims involving *bottom-level* interactions, which purports to solve the regress, or bottoming-out, problem. Here Glennan favours a *dispositionalist* view, by suggesting—although not articulating—the view that singularism leads to interpret bottom-level interactions in terms of the manifestations of 'powers':

> [the singular determination view] holds that there are genuine interactions between parts at the bottom of the mechanistic hierarchy, but that these parts are not governed by laws. In calling these interactions genuine, I am suggesting that the relationship is a modal one. We can express the modality of the relationship counterfactually: When a change in a produces a change in b, it follows (with the usual caveats about overdetermination, etc.)

> that if a had not changed, b would not have changed. But the counterfactual locution should be understood not as a claim about non-actual worlds, but a claim about the determining power of a in this world (Glennan, 2011, p. 812).

Now, if Glennan wants to illuminate the meaning of causal claims by reference to their truth-makers, he must make clear the relation between the two views above.[73] The following options are open to him.

He could embrace a *pluralist* view: processes, however defined, provide truth-makers for claims involving higher-level causal relations, whereas dispositions provide truth-makers for bottom-level causal interactions. However, there are reasons why it seems implausible to read the mechanist account in a pluralist way. To begin with, 'process' and 'power' fall in the same conceptual category, namely 'production'. If production is metaphysically fundamental, as the mechanist claims, it should apply *across* levels, and to *all* mechanisms. Also, consider the case of two-part systems and causal claims describing interactions between their parts. Either one says these are not mechanisms (§5.3) or, since for any causal claim there is *one* truth-maker not two, there must be some connection between powers and processes. What is this connection? In any case, the account is incomplete.

An alternative to pluralism is to develop the mechanistic account into one or the other *monistic* view, by either reading the thesis on powers as a thesis on processes, in which case one gets a *process-based* account, or the other way round, so as to get a *power-based* account.

Building on Glennan's latest view on powers, an obvious alternative for us to explore is whether a *dispositionalist* story can help make sense of Glennan's general idea that mechanisms are systems/processes ultimately grounded in the dispositions of their parts, and to develop this idea into an account of causality in complex systems. For one thing, in fact, it would be nice to have a coherent story that explains the relation between powers/dispositions on the one hand, and processes on the other. For another, the existence of such a relation could also explain the nature of fundamental level interactions and allow for the extension of the notion of mechanism to two-part systems. Just to recall the problem: Why aren't 'simple' two-part systems whose parts

[73] Glennan maintains (personal communication) that his views on powers and processes are not in opposition, but rather complement each other: at the bottom of the hierarchy there is a brute set of powers, whereas further up the hierarchy and with further extension in time between events, there are mechanically explicable processes. Yet, the pluralism suggested by this view does not sit well with the mechanistic project, as I explain below.

interact locally mechanisms? Even if no explanation by further decomposition is possible, if the notion of production is metaphysically fundamental, shouldn't it apply across levels, and to fundamental level interactions as well? At least, this the kind of take one would expect from a dispositionalist. In line with Glennan's recent view that mechanisms and interactions are underpinned by their parts' dispositions, I will first try to re-interpret Glennan's theory in a dispositionalist framework and then test the re-interpreted theory against examples of causal relations in complex systems.

Before, however, let me briefly digress to show what Glennan's mechanisms are *not* to be identified with, viz. Salmon-Dowe processes—which will prove that one of the two alternatives, viz. the reduction of powers to processes, is not viable.

5.4.4 Mechanisms are not SD processes

One may try to reduce powers to processes, so as to analyse mechanisms in terms of processes. The success of this move requires that one specify what counts as a *causal* process, as opposed to a *non-causal* process, otherwise one has not even a metaphysical account of *non*-bottom-level causation. In order to do this, one has to rely on some available account of causal processes or develop an alternative. As I argue below, our current understanding of causal processes, as described in Salmon-Dowe process theory (henceforth, SD), does not fit complex systems mechanisms. To be fair to Glennan, it must be said that he himself at various times has pointed out that his account of causality is different from SD (Glennan, 2002, 2011). It is worth stressing, however, the reasons *why* complex systems mechanisms are not SD processes.

In short, SD is based on the following definitions (see Dowe, 1995, p. 323)[74]:

[SD$_1$] A causal *interaction* is an intersection of world lines which involves exchange of a conserved quantity.

[SD$_2$] A causal *process* is a world line of an object which possesses a conserved quantity.

The details of the account do not matter for the present purpose. What is

[74] Salmon's version (see Salmon, 1997, pp. 462, 468), which relies on the notion of 'transmission' rather than 'conservation' of quantities, is slightly different.

important is that SD is meant to apply to physics, and to whatever quantities physics says are conserved (e.g., momentum, mass-energy, charge). Since all levels of complexity for Glennan arise out of bottom physical interactions among fundamental particles, one may want to say that Glennan's mechanisms just are SD processes and interactions. The problem is that this move wouldn't do in complex systems.

To begin with, Glennan himself discards the hypothesis that complex systems mechanisms may just be SD processes, based on the following reasoning. A SD mechanism is a causal 'structure', or 'nexus', a sort of web of processes and interactions. Since in SD an 'object' is a 'process', one could (mistakenly) think that the marriage between Glennan and SD works as follows: Glennan's parts are objects; objects are SD causal processes; and the interactions between these parts are intersections in causal processes that introduce changes to the persistent structure of these processes, that is, changes to the properties of the parts (see Glennan, 2002, p. S346). However, Glennan rejects this hypothesis on the ground that 'Glennan-objects' are not 'SD-objects': an object for Glennan is a stable configuration of parts, whereas an SD-object need not have parts and can be an 'ephemeral' process. For instance, take the case where throwing a ball causes a window's breaking. Between the two events there is a process (the motion of the ball) which then interacts with another (the window at rest) so that mass-energy and momentum are globally conserved. But the ball-plus-window complex is not a stable *system* such as a cell or the market. In Glennan's words:

> The difference between the process theory and the mechanical theory lies in their rather different conceptions of what a mechanism is. For the process theorist, a mechanism just is a process of the sort described by their theory. To the mechanical theorist, however, a mechanism is a system (Glennan, 2011, p. 798).

In this sense, the SD notion of mechanism is *too permissive*. Therefore, we should not identify 'SD-processes' with 'Glennan-objects', or systems.

But there are further reasons why the marriage would be inappropriate. In another sense, in fact, the SD notion of mechanism is also *too strict*, as many *prima facie* complex systems would not count as mechanisms.

First, the identity conditions of a complex system need not depend on conservation of quantities, which is what defines an object in the SD sense. The SD notion of mechanism is of narrower applicability than Glennan's

(it only applies to physics). The requirement that an object be identified by conservation of quantities such as mass-energy is not met by complex systems, which maintain their identity through time in spite of, e.g., their growing or shrinking (e.g.: cells grow; agents enter and leave the market). Nor would considering features of the environment (e.g., resources, waste products) as 'parts' of the mechanism itself serve to re-establish the identity between Glennan- and SD-objects. On the one hand, (to a large extent, at least) a complex system is such-and-such an object *in spite*—rather than *in virtue of*—its surroundings. True, the system needs embedding in the environment to survive. However, it does survive across a large range of environmental conditions which, intuitively, do not all need mentioning for the identification of the system itself. On the other hand, even if (certain/all) features of the environment were included in the description of the mechanism, this would not by itself capture the ability of the system to change and adapt along with changes in such features and at the same time remain the 'same' system. Unlike SD objects, whose identity is fixed by the quantity being conserved, the boundaries between a complex system and the environment would need re-tracing along the evolution of the system-plus-environment complex.

Secondly, 'Glennan-interactions' are not 'SD-interactions': for one thing, Glennan's interactions are anything which can be characterised by direct, invariant, change-relating generalisations (a general, tolerant, counterfactual criterion), whereas SD's are exchanges of conserved quantities governed by conservation laws (a much narrower, physical, non-counterfactual criterion); for another, complex systems interactions are continuous changes in properties of parts that affect each other, whereas SD's are more like discrete spatiotemporal intersections—a 'nexus'—of processes.

For all the above reasons, a marriage between the mechanistic account and SD would not suit complex systems. Let us to turn, then, to evaluate whether the marriage between the mechanistic account and dispositionalism would be more successful.

5.5 A dispositionalist route to causality in complex systems?

An alternative to a process-based mechanistic account is to cash out mechanisms in dispositionalist terms. As we shall see, however, recent versions of causal dispositionalism seem unable to illuminate the meaning of causal

claims in complex systems. These considerations suggest that the dispositionalist project is not as promising as it might at first appear.

5.5.1 When are capacities efficacious?

To understand the job that dispositions should do in the mechanistic account, it is instructive to start from a discussion of the view that Glennan himself expresses in an earlier paper, namely Glennan (1997). There, Glennan tries to develop an account of capacities (i.e., a sort of dispositional properties) alternative to Cartwright (1989)'s and compatible with his mechanistic account. Glennan claims that capacities cannot be characterised as properties that operate *across contexts*, along the lines of Cartwright's CC condition (§4.3.1). This, as Cartwright herself acknowledged, does not square with the fact that, due to the existence of dual and interacting capacities, some capacities are not invariably connected to their manifestation. Glennan's diagnosis is that, to do justice to the primacy of singular facts over universal facts—to which both Cartwright and Glennan subscribe—capacities should not be characterised in terms of probabilistic relations among *classes* of events involving *populations*. Instead, they should be understood as properties of *individuals*, whose identity is (with the exception of fundamental capacities) *mechanically explicable* (see Glennan, 1997, p. 617). On this view, capacities are not new metaphysical kinds, but properties that individuals possess in virtue of their structure, and which are effective when their constitutive mechanisms are.

However, there are two problems with this understanding of capacities. First, and more obviously, fundamental capacities have no place in Glennan's mechanistic account, so one cannot explicate causal relations taking place at the bottom of the mechanistic hierarchy in mechanistic terms. If structural capacities are accorded no special (metaphysical?) role, and 'inherit' their powers from the arrangements of the individuals' parts, the mechanistic account is still vulnerable to the regress objection. Secondly, Glennan's account of structural capacities is not satisfying either. In fact, this view does not say what makes for the manifestation of one capacity on one occasion but not on another, even if the underlying structure is the same on both occasions.[75]

[75]Notice that a mechanistic account of explanation need not address this issue: for the purpose of explanation, it is usually enough to say that the phenomenon is—*ceteris paribus*—the result of a given collection of parts being arranged in such-and-such a way: the parts operate whenever they are arranged in the proper way. In contrast, a dispositionalist account of causality, whether mechanistic or not, must say what determines the manifestation of the disposition in order to provide the truth maker of the causal relation.

One would expect that the mechanistic account be informative on this issue, at least outside the domain of brute interactions. Following Cartwright, Glennan states that a capacity will manifest itself depending on how it interacts with other capacities. However, Glennan has no account of interactions tailored to complex systems.[76] As a consequence, he has no account of what makes capacities efficacious, i.e., of their relevance—hence, the relevance of mechanisms—to causality: Why does a causal relation obtain *in virtue of* a mechanism on some occasions, but it does not obtain on other occasions, *in spite of* the presence of the same mechanism? To solve the problem, one must modify the account of dispositions somewhere.

Recently, Cartwright has tried to make sense of the universality of capacities whilst maintaining the primacy of singular causal facts, by making the link between capacities and their exercise "analytic" (Cartwright, 2007a, p. 20). She offers a twofold interpretation of causal claims as either capacity claims, expressing facts about potentials for exercise (e.g.: 'Smoking has the capacity to cause cancer'; 'Aspirins have the capacity to relieve headaches') or causal laws, expressing facts about manifestations (e.g.: 'Smoking causes lung cancer'; 'Aspirins relieve headaches'). A true capacity claim is true *across contexts*, in the sense that it refers to the link between the capacity and its "potential for exercise", not between the capacity and its manifestation (cf. Cartwright and Efstathiou, 2007). A true causal law, instead, is true only *ceteris paribus*, as it describes the operation of mechanisms, or 'nomological machines', whose operation tends to produce regularities, but may be hampered, neutralised, etc. A nomological machine is defined as

> a fixed (enough) arrangement of components, or factors, with stable (enough) capacities that in the right sort of stable (enough) environment will, with repeated operation, give rise to the kind of regular behaviour that we represent in our scientific laws (Cartwright, 1999, p. 50).

So, causal claims are true in virtue of the repeated operation of mechanisms, which in turn depends on the operation and manifestation of the capacities of their components. This view commits one to a "metaphysically heavy" position, viz. a trichotomous distinction between presence-exercise-manifestation

[76] Glennan says that a characterisation of interactions needs to make reference to physical theory, and mentions approvingly Salmon's treatment of interactions (Glennan, 1997, p. 612, fn. 3). However, it is clear that this move won't do, because of the incompatibility between 'SD interactions' and 'complex systems interactions' (§5.4.3).

(see Cartwright, 2007a, pp. 19-21, 24): whenever a capacity is present and/or properly triggered, it will operate (the presence-exercise link is necessary); however, a capacity may or may not manifest itself, depending on how it interacts with other capacities (the presence-manifestation link is contingent). The pay-off of this position should be that, by postulating capacities, one can more easily bridge testing and use of causal relations, viz. *explain* the evidence (i.e., the obtaining or non-obtaining of phenomena) and *act* upon it (i.e., devise 'effective strategies' that bring about changes by triggering or inhibiting the individuals' capacities). But there are reasons why an appeal to capacities is not very illuminating in the case of complex systems.

Establishing what a capacity does comes in two steps. First, one calculates how the capacity works in isolation, e.g., by eliminating or calculating away everything that can interfere with the production of the effect in experimental conditions where we know which factors are present and how they operate (see Cartwright, 2007a, pp. 2-3). In the absence of the possibility of controlled experiments one can rely on, e.g., a well-conducted RCT. In this way, one establishes a capacity claim. However, making use of a capacity claim requires knowledge that the corresponding causal law is true. This second step requires calculating the contribution to the effect in target, non-experimental conditions. To this end, knowing that the effect *may* occur is not enough, one should also know that it *will* occur, and will do so with some particular *strength*. This presupposes knowledge of how the other capacities operate and the way they interact with one another and with the one whose contribution we want to calculate. Sometimes, there are nice rules that allow one to infer true causal claims from capacity claims. One example are the textbook problems of classical mechanics where the systematic difference produced by a cause (its contribution) can be calculated by means of an additive rule. Unfortunately, the case of complex systems is not one of these lucky cases. Due to nonlinear interactions, sensitivity to initial conditions and openness to the environment, knowledge of the presence of a capacity is not so informative as to how it will interact with other (possibly many) capacities present, which capacity will manifest itself, and with what strength it will contribute to the effect.

So, one may agree with Cartwright that capacities operate across contexts. However, anything has the capacity, in principle, to bring about, or prevent, almost anything else, *in suitable conditions*. Thus, the informativeness of the capacity claim is parasitic on the specification and generality of such

conditions. The more complicated the conditions, the less informative the capacity claim. As a result, in complex systems the appeal to capacities proves less illuminating than in other, more convenient cases.

Notice that the lack of informativeness of dispositions in cases of causation in complex systems is not limited to Cartwright's account of the relation between causality and capacities, but applies to other dispositionalist accounts of causality, too. I'll only mention one such account, namely Chakravartty (2007)'s.

5.5.2 What fixes the identity of the truth-makers?

Chakravartty (2007)'s dispositionalist account purports to elucidate the connections between (causal) processes, dispositions, events and properties. As such, it seems to offer the resources to develop Glennan's mechanistic account into a power-based mechanistic account. Yet, as I argue below, one cannot adopt Chakravartty's notion of disposition to fix the problems with the mechanistic account.

In Chakravartty's view, properties are primitive, and individuals (objects, events and processes) are derivative from them. Individuals are bundles of properties, some bundles being more stable, other bundles being less stable (see Chakravartty, 2007, chap. 6). For instance, having molecular structure H_2O and being drinkable are (quite) stably related. Instead, being a Bcl2 protein and promoting a caspase cascade are less stably related, and so are being an increasing time series of an asset's price and being a determinant of a crash. In particular, "[a] causal property is a property 'conferring' to particulars that have it dispositions to behave in certain ways when in the presence or absence of other particulars with causal properties of their own" (Chakravartty, 2007, p. 108). The identity conditions of a causal property are fixed by what Chakravartty calls the "dispositionalist identity thesis" (DIT): "what makes a causal property the property that it is are the dispositions it confers to the objects that have it" (Chakravartty, 2007, p. 129).

What gives dispositions causal efficacy? The '*de re* necessity' of causal connections that follows from DIT (see Chakravartty, 2007, pp. 113-114, 121). DIT entails a holistic, mutual determination of causal laws and identity of causal properties, which Chakravartty dubs "ontological circularity":

> not only (...) causal laws comprise relations between causal properties, but also (...) knowing such laws allows one to distinguish

> and identify properties as well. A causal property can be identified as the property that it is in virtue of its relations to other properties. The conjunction of all causal laws thus specifies the nature of all causal properties (Chakravartty, 2007, p. 123; see also p. 140).

Let us imagine that at a certain world a certain number of properties exist. By being there in place, they all at once co-determine their own *identity*, that is, the dispositions for behaviour they confer to the individuals that possess them. Also, by being there in place, the properties co-determine the *relations* they have with one another. Such relations are of 'natural necessity': of 'natural', or *de re*—as opposed to 'metaphysical'—necessity, because the necessity need not hold across possible worlds; and yet of 'necessity', because at this world the relation between any two given properties is fixed by DIT. Take, for instance, a given distribution of, e.g, Apaf1, pro- and anti-apoptotic proteins, procaspases 9 and 3, Smac, XIAP, etc. This makes it possible that the mitochondrion releases cytochrome *c*, cytochrome *c* binds to Apaf1 thereby forming the apoptosome, the apoptosome activates procaspase 9, Smac inhibits XIAP, XIAP inhibits Casp9, Casp9 activates procaspase 3, etc. At the same time, in turn, the latter relations fix the identity of the properties, that is, determine the disposition(s) for behaviour they confer to the individuals that possess them. For instance, pro- and anti-apoptotic proteins are classified as belonging to the same class in virtue of their function of regulating apoptosis via the regulation of the release of cytochrome *c*.

Notice that the necessary relation between the properties does not entail that the relations *have to* give rise to certain (relations between) manifestations, or events. Whether a given relation between event c and event e obtains will depend not just on properties C and E of the events in questions, but also on other properties of (the events that we label) c and e, on what is between c and e, as well as what is at other places. This follows directly from the holism with which the properties co-determine each other.

This dispositionalist view has both advantages and disadvantages with respect to the goal of characterising causality in complex systems.

One advantage of endorsing this picture is that it gives a more coherent story on the relations between powers, processes, events and objects, hence a possible ontological foundation to the mechanistic account. Processes are linked to dispositions (or powers) by being the (continuous) manifestation of

the exercise of such dispositions. Manifestations, in turn, whether events or processes, are grounded in the interactions among properties. The interactions among the properties determine, at once, all particulars, i.e., objects, processes and events. Causal phenomena are the result of continuous processes of interaction among particulars (i.e., objects, processes and events) with causal properties (Chakravartty, 2007, pp. 108-109). Causation itself is a (continuous) relation taking place between properties (or property instances), and only derivatively between events. Relations between (discrete) events are just epiphenomena, which we pick out of the field of "continuous processes of interaction" for reasons of salience. In particular, the realist's talk of events as *relata*, although convenient, is only elliptical for descriptions of "aspects" of such processes.

This account can also tell a story on how a mechanism can be both an object *and* a process: it is an object insofar as a given bundle of properties is relatively stable; it is a process insofar as one or more of these properties change through time, either in virtue of the interactions among the parts of the system, or in virtue of influences coming from the environment.[77]

A final advantage is that it becomes (more) plausible to treat two-part systems as genuine mechanisms. Given that the boundaries of the systems are to an extent conventional/artificial, strictly speaking no system has three parts any more than it has two or ten. What matters is the strength of the couplings. A system will be identified by strong(-er) couplings among its 'parts' and weak(-er) couplings with the 'environment'. In general, whether we can call a given system a mechanism will depend on how flexible is our definition of mechanism.

However, there are also disadvantages in embracing this picture, having to do with the holistic baggage that comes with Chakravartty's dispositionalism. What is the truthmaker of a causal claim? According to Chakravartty,

> it is a consequence of DIT that networks of causal properties have a holistic nature. This furnishes a more radical solution to the problem of truthmaking than it is generally appreciated. The existence of any one causal property is a sufficient truthmaker for counterfactuals about all possible relations applicable to the world

[77] Notice that, *contra* Glennan, on this picture function is as important as structure: it is both true that a property's identity determines (structurally) the causal relations that property is involved in and that the causal relations determine (functionally) the property's identity.

> in which that property is found (Chakravartty, 2007, p. 146).

Chakravartty's dispositional holism seems incompatible with Glennan's locality requirement. If Chakravartty is right, it is not possible to identify the truth-maker of *any one* causal relation without at the same time having to mention, strictly speaking, *all other* causal relations. If so, no causal relation is strictly speaking local: whether a causal relation between any two events holds depends on the complex result of all the causal relations present at a given world. This will depend on whether the—many—processes that enable the obtaining of the relation are not hampered by the—*very* many—processes that can in principle prevent it. Relatedly, Chakravartty's dispositionalism is silent on what (local) factors are more relevant than others to the truth of causal claims.

When evaluating what features of the mechanism are responsible for the causal relation, what one is normally interested in are not the parts' dispositions themselves but rather the way in which they give rise to their *manifestations*. Moreover, one is normally interested in the *local* not global determinants of these manifestations. Finally, one is interested in the *thread* between one event and another. Does buying this stock cause a bubble? What is the truth-maker of 'Switching of fundamentalists into chartists caused the bubble'? What explains a specific event (e.g., a bubble) or a general pattern (e.g., fat tails, volatility clustering, volatility persistence)? Dispositionalism should prove, if not sufficient, at least particularly useful to answer these questions. Unfortunately, there is only so much that Chakravartty can say about this:

> what one is often most interested in are the ways in which the states of objects evolve. States change, but explaining precisely how such change occurs is something that one can only say so much about. This is to concede Hume's point that ultimately one has nothing like a "picture" of what is happening when one thing brings about another beyond that which is observable or, one might add, detectable. That is why the demand for a causal mechanism cannot be fully satisfied (Chakravartty, 2007, pp. 111-112).

However, as I argue in §5.5.3, in complex systems not only this demand cannot be *fully* satisfied by an appeal to dispositions; it cannot be satisfied *in important respects*.

5.5.3 Dispositions are not enough

There are general reasons why the above dispositionalist strategies fail in fixing the mechanistic account and in delivering a better understanding of causality in complex systems.

First, dispositions are insufficient for identifying the truth-makers of causal claims. For instance, one cannot infer from DIT *alone* the *direction* of the causal relation. Once all properties are in place, global *de re* necessity (in Chakravartty's sense) among any two of them is holistically fixed. Yet, there is no clear sense in which DIT can deliver a verdict on the direction of the relation between particular events which emerges out of the continuous processes of interaction among the properties.[78] One could argue that there is an *intrinsic* directionality between properties which stand in a causal relation. Intuitively, smoking causes cancer, not vice versa, irrespective of other relations. However, holism challenges this—very intuitive—idea: change the context in the appropriate way and the directionality of the relation changes as well. Due to phenomena of 'reverse causation', at different times in different contexts something can both cause and be caused by something else. For instance, p53, responsible for promoting the synthesis of pro-apoptotic proteins, also promotes synthesis of Mdm2. The latter protein, in turn, 'tags' p53 and makes it digestible by proteasomes. By means of this negative feedback loop, p53 regulates itself, thereby regulating apoptosis in general. So, depending on the context, it is both true that a change in p53 level causes a change in Mdm2 level and that—vice versa—a change in Mdm2 level causes a change in p53 level. Analogously, in the asset pricing mechanism, a bubble can cause one to buy, which in turn, can reinforce the bubble. So, knowledge of dispositions is insufficient to determine whether 'this' causes 'that'.

Secondly, dispositions may be also unnecessary, if it were possible to 'bypass' them and always infer occurrent (i.e., non-dispositional) properties from other occurrent properties without postulating, and referring to, underlying dispositions.[79] At this point, the dispositionalist usually objects that dispo-

[78] Usually, one decides on the direction of the causal relation with the aid of pragmatic considerations such as salience or manipulation, not in virtue of knowledge of dispositions alone. However, since Chakravartty is concerned with ontic not pragmatic issues, the latter have no definite place in his account of causality. All the more reason to believe that dispositions *alone* are insufficient.

[79] This leads to envisaging dispositions as a sort of 'inference tickets'. Although Chakravartty does contemplate this possibility, he then discards it because incompatible with causal realism (see Chakravartty, 2007, p. 124). Interpreting causes as inference tickets goes towards the approach I develop in chapter 8 (although my proposal has no eliminativist pre-

sitions are still useful, on the ground that only dispositions can give counterfactual claims their modal strength. For instance, the truth of 'If XIAP level were to change, Casp3 level would change' *must* depend on something like the disposition of XIAP to inhibit/promote Casp3. The problem is that the dispositions of individual parts in complex systems shed only little light on the obtaining of causal relations. Let me explain.

In general, due to the many couplings inside and outside complex systems, knowledge of dispositions of the individual parts is not as informative as knowledge of geometrico-structural features of the whole—which need not be stably attached to specific individuals (cf. Goldstein (1996), Smith (1998, chap. 7) and Kuhlmann (2011)). For instance, although the dispositions of, say, XIAP, Casp3 and Casp9 are somehow important, they are also context-dependent. XIAP's binding to Casp3—which is something that would 'normally' *prevent* apoptosis—can also *promote* apoptosis. Structural features of the context (e.g., relative concentration of the reactants, their position, etc.) may be more informative than the reactants' dispositions. This context sensitivity is even more obvious in the case of asset pricing. Here, knowledge that the disposition of an individual trader to buy a stock can *promote* a bubble tells very little, because depending on the context it may also have the opposite disposition, viz. it can *prevent* the bubble. More informative are system parameters (e.g. thresholds for the normal-to-chaotic transition, rate of change in the traders' attitude and/or trading strategies, etc.).

Furthermore, since identity and function of individual parts can change during the process, dispositions need not be stably attached to them. Fundamentalists may become chartists, optimists may become pessimists. Explanation by reference to the chartist behaviour of individual agents misses the important fact that the agents' chartist disposition may depend on contextual reasons, not on their 'intrinsic', chartist nature. Analogously, in the apoptosis mechanism caspases are synthesised as inactive and become active by proteolitic cleavage. One may say that caspases just are procaspases *disposed* to become active. However, explanations in terms of procaspases' disposition to become active are limited. In fact, in certain contexts XIAP prevents apoptosis and procaspases become caspases, in other contexts XIAP promotes apoptosis and procaspases fail to become caspases—without either behaviour being explainable just in terms of XIAP and procaspases' dispositions.

tension). Notice, however, that whether inferentialism commits one to causal anti-realism is an open issue, which may depend on the way inferentialism itself is interpreted.

To sum up, although it may well be true that C 'disposes towards' E, in complex systems whether C 'causes' E hinges more on the context than on the disposition. Depending on the context, C may either promote or prevent E, and—vice versa—E may either promote or prevent C. Not much of an indication as to what causes what. What we have, then, is a superabundance of 'analytic' facts about dispositions, using Cartwright's jargon, and too little knowledge about their possible manifestations (i.e., knowledge of the form: in context X, disposition C would bring about effect E) for the dispositionalist account to be really informative about the meaning of causal claims.

5.6 Mechanistic models and causality

In §4.5, I expressed reservations towards the idea that the only viable theories of meaning are in terms of truth conditions. Here, I want to make a parallel point with reference to the view that the only viable interpretation of the way causal models represent is based on the *similarity* or *isomorphism* between models and their target systems—viz. the so-called 'semantic view' (van Fraassen, 1980; Giere, 1988; Suppe, 1989). Although arguing against the semantic view goes beyond the scope of this work[80], I do want to stress here that it is such a view that implicitly motivates the attempt of certain causal realists to account for the meaning of causal claims by reference to their truth-makers.

As Chakravartty notices, the idea that a particular understanding of how theories and models represent may justify or 'facilitate' realism about the entities such theories and models talk about has recently gained some currency among certain proponents of the semantic view. However, Chakravartty argues, this idea is a non-starter (see Chakravartty, 2007, part III). One may extend Chakravartty's considerations from the way the proponents of the semantic view appeal to the success of models for justifying the existence of theoretical *entities* to the attempts to explicate the success of causal models by reference to causal *relations*. My point is that the idea that the meaning of causal claims derivable from models of complex systems is best explained in terms of the truth-makers that ground the success of such claims follows directly from the more general idea that successful representation is constituted by a correspondence or similarity relation between a class of relations in the model and a class of worldly states of affairs. And since we don't need

[80]For recent criticisms of the semantic view, see, e.g., Suárez (2003) and Frigg (2006).

to buy the latter idea, we don't need to buy the former either.

Interestingly enough, the attitude described by Chakravartty can be ascribed to Glennan, too. Parallel to his account of causation, Glennan (2000, 2005) has developed an account of 'mechanical' models, which explicitly relies on the semantic view. Ultimately, on this account the notion of mechanism as the substrate of causal relations is explicated in terms of the conditions under which the model with respect to which the causal claim is formulated is a faithful representation of its target system. For Glennan, whether the model represents a mechanism depends on whether variables and functional relations in the state space are interpretable in terms of, respectively, property of parts and relations among them (see Glennan, 2005, pp. 447-448).

The thing is that, since this representationalist interpretation of mechanisms depends directly on Glennan's conceptual explication of 'mechanism', it is not very enlightening as regards the meaning of causal claims in complex systems. Luckily, the semantic view is not the only game in town to explain scientific representation. Inferentialism may constitute a viable alternative. Although inferentialism is primarily a theory of meaning, there is a growing interest with regard to the applicability of inferentialism to issues such as scientific representation (Suárez, 2004; de Donato Rodríguez and Zamora Bonilla, 2009a) and explanation (de Donato Rodríguez and Zamora Bonilla, 2009b). This motivates an attempt at a reinterpretation of the issue of the grounding of the truth and explanatory power of a causal claim in terms of the conditions under which a causal model *represents* (inferentially) its target, which ultimately depends on whether the claims that the model licenses are correctly assertible or not.

Conclusion

Glennan's mechanistic account of causality is explicitly meant to fit complex systems. However, it is threatened by problems of circularity and regress, which it does not fully solve. Such problems arise because the account contains ambiguities as regards the notions of mechanism and interaction, which are meant to help account for the truth-makers and the explanatory power of causal relations. Various routes to make the account coherent are blocked, e.g., the manipulationist route and the SD process route. The ambiguities that vitiate Glennan's account may be eliminated by reference to a dispositionalist metaphysics, towards which he himself has lately gestured. However,

reference to dispositions of specific parts with stable identities and functions for providing truthmakers of causal relations in complex systems and explaining complex systems' behaviour proves less appropriate vis-à-vis reference to structural features of the arrangement. In the next chapter, I will argue that pluralist accounts of causality give us no good insight into the meaning of causal claims in complex systems, and that inferentialism need not entail any strong pluralism on the concept of causality.

Chapter 6

Pluralist Accounts of Causality

As a response to the failure of monistic accounts, the view that causality is a diverse notion is becoming more and more popular. After describing the spectrum of pluralist positions (§6.1), I present the monist's main objection against the pluralist (§6.2): How can we reconcile the idea that causation is diverse with the fact that the label "causal" is used to denote all these relations? Is there (not) something they all share? I argue that both 'determinate' pluralism (§6.3) and 'indeterminate' pluralism (§6.4) are unable to answer these questions satisfactorily. I then move on to address *semantic* pluralism, a position recently advocated by (Reiss, 2011), in opposition to a merely *epistemic* pluralism (Williamson, 2006). Reiss explicitly appeals to inferentialism to justify his semantic pluralism. After sketching the inferentialist approach to semantics (§6.5), I interpret the debate between evidential and semantic pluralist in inferentialist terms (§6.6). Finally, I argue that inferentialism *can* account for the monist challenge (§6.7). In particular, I argue that inferentialism can explain why the concept of causality is at once monistic in *one* sense, and pluralistic in *another* sense. As such, inferentialism offers itself as an ideal framework for discussing both the prospects of conceptual monism and the meaning of causal claims in specific areas of inquiry, complex systems sciences included.[81]

6.1 A plurality of pluralisms

The label "causal pluralism" has been recently attached to a variety of positions, all sharing the idea that there are distinct kinds of causal relations, and no single feature that makes all of them causal—from which it follows that no monistic account is possible. In Nancy Cartwright's words:

[81] §6.1–§6.2, and §6.5–§6.7 are reproduced, with minor modifications, from (Casini, 2012).

> Under the influence of Hume and Kant we think of causation as a single monolithic concept. But that is a mistake. The problem is not that there are no such things as causal laws; the world is rife with them. The problem is rather that there is no single thing of much detail that they all have in common, something they share that makes them all causal laws (Cartwright, 2004, pp. 813-814).

But underlying this one, shared idea is a plurality of positions, which can be classified along *two* dimensions, viz. (i) determinate vs indeterminate pluralism (Williamson, 2006), and (ii) evidential vs semantic vs metaphysical pluralism (Reiss, 2011).

Causal pluralism is *determinate* if it maintains that there is a *finite* number of distinct (i.e., irreducible to one another) characterisations of 'causation', such as difference-making and production. It is *indeterminate* when it appeals to the (Anscombian) view that there is an *indefinite* number of notions of cause, viz. as many as there are causes, and that pushings, pullings, breakings, bindings, etc. are substantially different from one another and share no common truth-maker (Anscombe, 1971). According to the latter view, these various causings are best rendered by 'thick', or 'content-rich', causal verbs (Cartwright, 2004). This view I label "thick-concept view" (§6.3).

In the determinate camp are those (Hall, 2004; Longworth, 2010) who argue that "causes" has essentially disjunctive meaning, i.e., it either means x, or y, or... Counterexamples and objections, however, have been offered to undermine determinate articulations of pluralism (cf. Longworth (2006) on Hall (2004), and Cartwright (2010) on Longworth (2010)).

In the indeterminate camp, instead, the positions (Psillos, 2010; Cartwright, 2004; Reiss, 2011; Godfrey-Smith, 2009) range from holding that there is a substantial diversity among the token cases of causation to admitting the existence of a—more or less strong—family resemblance among them, which may amount to a sort of *weak*, or "nebulous" (Williamson, 2006, p. 74), *monism*. But the various facets of the indeterminate views are harder to disentangle.

In fact, orthogonally to the determinate vs indeterminate distinction, one may further distinguish between evidential, semantic (or conceptual) and metaphysical pluralism. An *evidential* pluralist (Russo, Williamson, Psillos) is pluralist in the minimal sense that he acknowledges that evidence of more than one kind contributes to establish a causal claim. A *metaphysical* pluralist

	evidential	**metaphysical**	**semantic**
determinate	Russo, Williamson	Hall, Longworth	Hall, Longworth
indeterminate	Psillos	Cartwright (I)	Reiss

Table 6.1: spectrum of pluralist positions

(Cartwright) maintains that the truth-makers of causal claims are different kinds of causings (e.g., although pushings and pullings are both causings, they also differ from one another). A *semantic* pluralist, finally, claims that there are various notions of cause—whether determinate (Hall, Longworth) or indeterminate (Reiss). Whilst metaphysical and evidential pluralism are in principle compatible with conceptual monism, semantic pluralism is clearly not (§6.6).

Table 6.1 shows how the various pluralist positions can be classified along these two axes. Notice that in this table Cartwright figures as a metaphysically pluralist, insofar as she (often) claims that causings are essentially different, and best characterisable by thick concepts. However, as I am going to explain, she seems to hold also a 'weakly-monist' view (table 6.2) which I label "inference view" (in short, INF), since it appeals to broadly inferentialist considerations. I clarify what INF consists in §6.2 and §6.5. For the moment, it is just important to stress that TC and INF are *distinct* and can be held *independently*. Since both Reiss (2011) and Williamson (2005, 2006), as it turns out, hold INF, a natural question to ask is whether, in virtue of holding such a view, one should be semantically or just evidentially pluralist. I address this question in §6.6 and §6.7.

6.2 The monist's challenge

Against all pluralisms—except the *evidential* one—the monist (Williamson, 2006; Russo and Williamson, 2007) maintains that, although causal claims may be supported by distinct evidential criteria, there is just *one* notion of cause. This position is in line with the epistemic view of causality (Williamson, 2005, 2006), according to which a causal relation is the inference relation drawn by an ideal, fully rational and informed agent.[82] Since the epistemic

[82]It is worth clarifying in what sense the epistemic view can both count as pluralist in one sense and monist in another: *contra* 'traditional' monistic analyses, the epistemic view is—minimally—pluralist as it doesn't erect any *test* condition to the status of *truth* condition; at the same time, it also maintains that there is something *all* causal relations have in common, namely their *essentially* being ideal inference relations—which *is* a thesis about

view purports to say what causality 'really' is, it is a metaphysical view. However, insofar as it defines the concept of causality in terms of inferences, it also counts as a position on the semantics of causal claims, namely a sort of inferentialism. Only the latter aspect, and not the details of the epistemic view, is relevant to the present discussion. In fact, I will only be concerned with the objection raised by the epistemic monist against pluralism, and in particular with its implications with regard to Reiss' *semantic* pluralism. Accordingly, my argument will only concern the semantics, *not* the metaphysics of causality.[83]

The monist objects to the pluralist that he cannot explain why, depending on the circumstances, he appeals to one *or* the other criterion (*contra* indeterminate pluralism), or to *several/all* criteria (*contra* determinate pluralism). Why in most cases do several, or even all, criteria apply equally well? And what principled reasoning, if any, is behind the choice of one criterion rather than another in a given context? If one believes these are philosophically interesting questions, then one will demand that a theory of causality provide answers to them. Notice that whether these questions deserve a philosophical—rather than, say, a historical—answer is a matter of controversy.[84] In §6.7, I endeavour to show that an interesting, philosophical story *can* be given, viz. an inferentialist story.

According to the monist, one will judge pluralist accounts as more or less adequate or desirable depending on how well they approximate the informativeness of a monistic account. At the uninformative end of the spectrum lies indeterminate—or "nebulous", as Williamson also dubs it—pluralism:

> [The] nebulous variety of pluralism is a last resort. If one can't say much about the number and kinds of notions of cause then one can't say much about causality at all; this stance should only be adopted if there is no viable alternative (Williamson, 2006, p. 72)

truth conditions.

[83] This is not to say that inferentialism is incompatible with giving causal talk a referential value, or being 'realist' about causality. Only this is not my concern here (for more on this, see §8.1.2). As regards the point I wish to make in this chapter, one may be anything from eliminativist (like, e.g., Psillos, 2010) to Anscombian pluralist (à la Cartwright).

[84] Psillos (2010) and Reiss (2011) have on different grounds claimed that there is no deep answer to be given, since there is no deep fact of the matter behind the use of the common label 'causal'. In particular, Psillos maintains that the answer cannot be deep because no deep metaphysical story can be given—there isn't any "one single, unique fully definite, etc. truth-maker for all causal truths". One may agree on this, and still disagree that the answer cannot be deep, provided one does not require that 'deep' answers be in terms of *metaphysical* essences, or truth-makers.

The family-resemblance view associated with Cartwright's TC clearly counts as nebulous in this sense. At the more informative end are monist positions. One such position is, as I said, epistemic monism. Another, as I argue in §6.7, is the version of causal inferentialism I sketch in §6.6, which envisages causality as one, vague cluster concept.

Curiously enough, Cartwright herself points to the possibility to interpret causal claims in inferentialist terms (Cartwright, 2007b, p. 46): what makes the plurality of different kinds of causal relations 'causal' is not some special relation in the world but some unified features of the representations themselves, where 'representation' is understood in *inferentialist* terms, i.e., scientific theories and models represent in virtue not of some alleged similarity or isomorphism they bear with the portion of reality they aim to represent, but in virtue of the inferences they license about it (Suárez, 2004). For Cartwright, this proposal is particularly appealing in the case of the representational meaning of causal claims, due to the traditional connection between the concept of causation and a particular kind of inferences, viz. inferences about the result of interventions. An inferentialist approach would then lead naturally to the project of 'making explicit' (using Brandom's jargon) the concept of causality in terms of inferential connections (see chapter 7).

That inferentialism *may* have monistic implications is suggested by Cartwright herself who, after pointing to the possibility to interpret causal claims as inference licenses (cf. Cartwright, 2007b, p. 46), states: "What we should be looking for is a theory of causality, in much the same way as we have a theory of the electron" (Cartwright, 2007b, p. 52). It is now commonplace that there are no necessary and sufficient conditions that *define* theoretical terms; still, we have theories that give *one* story about them. Cartwright seems to think—here, at least—that the same reasoning applies to causality. In particular, a theory of causality should be tied to the strategies to hunt causal relations on the one hand, and to the strategies to use them on the other (see Cartwright, 2007b, pp. 48-49).[85]

To the inferentialist camp belong, besides Cartwright, also Godfrey-Smith (2009) and Reiss (2011). Yet, they define their inferentialism differently—so that only Godfrey-Smith but not Reiss can be associated with the sort of weak, conceptual monism I am arguing for in this chapter (§6.7.2). Although

[85] At other times, however, Cartwright is skeptical about the chances to come up with such a 'theory' (cf. Cartwright, 2010, p. 327)—which seems the reason why she merely points to the inferentialist alternative, without exploring it.

	metaphysical	**semantic**
nebulous		Cartwright (II), Godfrey-Smith, Psillos
determinate	Williamson	Williamson

Table 6.2: spectrum of monist positions

all of them envisage causality as a cluster concept, they interpret the cluster differently. Table 6.2 classifies the various monistic positions.

Interestingly, what the advocates of both epistemic monism (Williamson, 2005, 2006) and inferentialism (Reiss, 2011) have in common is that they tend to characterise the cluster in terms of *inferential* relations (§6.6). Reiss, in particular, explicitly draws his pluralist conclusion from an inferentialist approach to the semantics of causal claims. So the question arises as to whether endorsing inferentialism need to commit one to semantic pluralism (Reiss) or not (Williamson). As I argue in §6.7, inferentialism need not entail a strong semantic pluralism.

I will now review the various pluralist positions and evaluate them according to how well they account for the monist's challenge, viz. why is the same label "causal" used to denote apparently different relations?

6.3 Determinate vs indeterminate pluralism

Let us consider determinate varieties of pluralism first. The best known pluralist account is the two-concept account offered by Hall (2004): C causes E iff (C produces E) *or* (E depends on C). This account regiments the dichotomy of intuitions which has been used to distinguish between the two broad categories of monistic accounts presented in chapters 4 and 5.

However, it has been suggested that there are cases that do not fall under either disjunct (Longworth, 2006, pp. 59-60). E.g.: A and B can prevent C, which would otherwise prevent E. In the circumstances, A prevents C, and E obtains. There's no mechanism involved here: it is an absence that causes E. And there's no dependence either: were it not for A, B would have prevented C, so E would have obtained anyway. The counterexample purports to prove that in cases like this our intuitions converge on the judgement that there can be causation without either production or dependence.[86] Rather than dismissing pluralism altogether, Longworth (2010, p. 314) suggests that the

[86]For a different counterexample, see Schaffer (2000).

analysis be replaced by a *longer* disjunction—which I will call the 'disjunctive-concept' analysis (DC). For instance, C causes E iff between C and E there is manipulability and probability raising, *or* locality and transference, *or* counterfactual dependence:

[DC] C causes E iff (i) INT & PR $\vee$ (ii) SD $\vee$ (iii) CD.

Notice that Longworth does not explicitly endorse *this* modified proposal, only suggests that some longer disjunction *or* other will fix the problem with Hall's two-concept analysis. For instance, DC does not account for cases involving neither dependence nor production, because it forgets about *other* disjuncts, e.g., 'C is morally *responsible* for E' (Longworth, 2006, S5). But the idea is that some disjunction exists such that for any given causal relation (i) the obtaining of one or the other disjunct is sufficient to make the relation causal (if the state of affairs involving C and E satisfies one disjunct, then there is causation between C and E) and (ii) the whole disjunction is necessary (for any causal relation, the features that make it causal must be listed, i.e., the disjunction is exhaustive).

Longworth then proposes to measure the causal content of a causal concept (its 'thickness') by reference to the number of disjuncts it entails: the more disjuncts, the greater the content. Take the claim 'The carburetor *feeds* gasoline and air to a car's engine'. Suppose its truth entails the truth of a corresponding locality-cum-transference claim. Then, 'feeds' has more causal content than 'causes' as appearing in 'The carburetor *causes* gasoline and air to be present in the engine', since 'feeds' entails information about locality-cum-transference, that is, a specific disjunct, whereas 'causes' entails no specific disjunct.

Against DC, Cartwright (2010) argues for an indeterminate pluralist view that revolves around two main ideas, namely TC and INF. First, the 'thin' verb 'causes' picks out a variety of *different kinds of relation* (each relation has "its own peculiar truth makers" (Cartwright, 2004, p. 817)) and has a variety of *different uses* (each use can be correct, depending on the context/purpose). The various kinds of relation and uses have little in common. They only bear a loose family resemblance to one another. In contrast, 'thick' causal verbs (e.g., 'pushes', 'pulls') are *more informative* than 'causes'. However, they are not reducible to traditional accounts of causality, or disjunctions of them.

They have some extra *causal* content which isn't captured by them. Secondly, the usefulness of the thin concept 'causes' derives not from its—allegedly unique—meaning but rather from the *formalisms* in which it figures. The assumptions which come with a formal system specify the conditions that (thick) causal laws must satisfy to obtain in some system. If the assumptions are satisfied, then they license a number of *inferences* of crucial importance for scientific practice. But there is no all-encompassing formalism that fits all systems (cf. Cartwright (2004, p. 818) and Cartwright (2007b, pp. 46-52).

It is worth stressing that the the two views can be held independently. In particular, it is possible to reject TC whilst endorsing INF, which is what I'll suggest one should do to give a satisfying account of the meaning of causal claims. As regards TC, this is best summarised in Cartwright's own words:

> All thick causal concepts imply 'cause'. They also imply a number of non-causal facts. But this does not mean that 'cause' + the non-causal claims + (perhaps) something else implies the thick concept. For instance, we can admit that *compressing* implies *causing* $+ x$, but that does not ensure that *causing* $+ x + y$ implies compressing for some non-circular y (Cartwright, 2004, p. 817).

Cartwright's point parallels a point made by the opponents of the so-called 'two-component analysis' of thick *ethical* concepts, e.g., 'cruel' (see Putnam, 2002, pp. 34-38). The advocate of the analysis maintains that 'This is cruel' is factorable into a descriptive component, e.g. 'This causes deep suffering', and an 'attitudinal' component, expressing some emotion/volition towards the fact in question, e.g. the speaker's disapproval of the cruel act. Since the attitudinal component isn't factual and is always attached to the descriptive component, the meaning of thick concepts is reducible to facts.

The opponent of the distinction agrees that ethical statements have (also) factual content, but argues that factual and ethical content are *entangled*: one cannot properly account for the meaning of 'cruel' without making use of 'cruel' or other *ethical* concepts. For instance, if 'causing deep suffering' is taken as having only factual value, a surgeon that cannot make use of anesthesia causes deep suffering but isn't cruel. Nor is cruelty reducible to any alleged factual component of 'causing suffering', e.g. 'causing pain', if 'causing suffering' is taken to have ethical value, too. Causing pain is not necessary for causing suffering—nor, *a fortiori*, for cruelty: a parent that prevents his son from fulfilling some talent causes suffering without causing pain. Thick ethical

concepts aren't reducible to factual descriptions. Analogously, Cartwright argues that thick causal concepts are not reducible to non-causal ones.

In the above quote, x stands for the non-causal facts (e.g., regularities among the variables in DAGs or structural equations), and y stands for the non-causal *differentia* which is supposed to make such facts causal (e.g., BNs' axioms, Woodward's invariance). If one indicates with "t" thick causal verbs and with "c" the thin 'causes', TC reads:

[TC] (i) $\forall t \exists x[t \rightarrow c \wedge x]$; (ii) $\neg \exists t \exists x \exists y[(t \rightarrow c \wedge x) \rightarrow (c \wedge x \wedge y \rightarrow t)]$.

Now, bearing in mind the above sketch of Longworth and Cartwright's accounts, let us come back to the debate between them. Longworth (2010, pp. 312-313) raises two objections against Cartwright's TC.[87] First, Cartwright would give no argument in support of the thesis that there is no non-causal x such that $c + x$ implies t. (Or, to stick to our previous formulation: there are no x and y such that $c + x + y$ implies t.) Second, she would not show that the extra (y-) content of t is actually extra *causal* content, rather than some *non-causal* 'nuance' (cf. Hitchcock, 2007). Such a nuance, admittedly, gets lost when the word "causes" is used instead, but because it is not a *causal* nuance it need not be part of a theory of *causation*. That is, the nuance need not belong to the analysis of the causal content of the thick description. So, we can replace thick concepts with DC. In the carburetor example, the nuance added by 'feeds' to the description of the locality-cum-transference fact is non-causal: (the causal content of) 'feeds' is exhausted by (the causal content of) 'causes'. If so, then it is false that 'causes' $+ x + y$ cannot imply 'feeds'. Furthermore, this opens the possibility that the causal content of 'feeds' is implied by locality-cum-transference facts *only*—y having the role of the non-causal nuance. As a result, DC would have the advantage over TC to reduce each case of causation to some clear set of conditions, with no extra, unspecified ("mysterious") causal content.

[87] Reiss (2011, fn. 3) takes pains to stress that Cartwright's pluralist view is best envisaged as a theory of *physical* causation not as a theory of the meaning of causal claims. Because of this, TC would not be subject to criticisms against pluralist theories of meaning. But there would be a tension, then, in Cartwright's position. If TC really is about physical causation only, then her criticism of DC would misplaced. If, instead, her theory is (also) about meaning of causal claims, as suggested by the above quote as well as by her debate with Longworth, then my own criticism against TC (§6.4) is well founded.

Remember that, for Cartwright, c and x are entangled. So, t's aren't reducible to any x alone. Why? She does not say explicitly. However, this could be, among other things, because the thick causal fact can obtain in a variety of ways, depending on the context. Just as there are contexts where sufference obtains in the absence of pain, there are also contexts where thick causal facts obtain in the absence of x. To all these contexts correspond different conditions (for which we may or may not be able to spell out formalisms). If so, then 'feeds' may not (always) imply x. It could imply *other* non-causal facts. So, the causal content of 'feeds'—generally conceived, not just as applied to the carburetor case—cannot be reduced to x. This seems right. Consider 'p53 *promotes* apoptosis'. This claim is true if—but arguably *not* only if—the following causal and non-causal facts obtain. The causal facts may be:

> p53 promotes synthesis of Mdm2 *and* p53 is phosphorylated or Mdm2 is *and* gene *p53* is not mutated *and*...

The non-causal facts, instead, may be encoded by, e.g., the following probabilistic relationship:

> Prob(pro-apoptotic proteins | Mdm2) > Prob(*no* pro-apoptotic proteins | Mdm2) *and*...

If such facts obtain, p53 promotes apoptosis. But is the following claim true, too?

> *If* p53 *promotes* apoptosis *then* Prob(pro-apoptotic proteins | Mdm2) > Prob(*no* pro-apoptotic proteins | Mdm2)

This claim may or may not be true, depending on the context. For instance, we know that p53 has not just the above mentioned, direct role in the intrinsic pathway, but also an indirect role in the extrinsic one, where it contributes to increase the cell's responsiveness to extracellular death ligands via the promotion of the expression of Fas-encoding genes. So, promotion of apoptosis due to p53 need not go through increase in Mdm2-induced synthesis of pro-apoptotic proteins; it can, instead, go through increase in Fas expression. So, 'promotes' has extra content with respect to the probabilistic relation obtaining in the intrinsic pathway.

However, is this enough to address Longworth's point? Notice that Longworth's claim was that t can be reduced to (i.e., entails) one *or more* disjuncts, the nuance being non-causal. This clearly allows for the possibility that t is realised in multiple ways. So, one may grant to Cartwright that the implication from 'causes' + x to 'promotes' doesn't hold in general. Consider 'promotes' and 'catalyses' as referring to the activity of a protein with respect to another protein. Both 'promotes' and 'catalyses' entail 'causes' and, arguably, probability raising. However, they do so in different ways. A protein *promotes* production of another protein by, say, binding to the promoters of the gene expressing it. Instead, a protein *catalyses* production of a protein by providing an alternative reaction pathway for the production of the protein. Facts about causation plus probability raising are not sufficient to decide whether promotion or catalysis takes place. So, the former notions do not exhaust the meaning of the latter. However, for Longworth's argument to go through it is enough that the implication from 'causes' to the thick concept holds true in each particular case *for some* x *or other*. Let us consider again the 'p53 causes apoptosis' example. If we focus on the single not the general case, then arguably 'p53 causes apoptosis' plus the probabilistic fact plus facts about the context do imply 'p53 promotes apoptosis', not 'p53 catalyses apoptosis'.

Also, by stressing that the obtaining of causation depends on the thick description satisfying the non-causal conditions, Cartwright seems to support—not contradict—Longworth's thesis that for each causing the job of specifying what makes the relation causal is done by the non-causal facts. If Longworth is right that the label "causes" attached to the re-description of the thick fact that satisfies x either doesn't add any content, being merely parasitic on the content of x, or adds something which is however mysterious, why shouldn't we drop it, and replace it with "clear and explicit" conditions (Longworth, 2010, p. 313)? Perhaps there really is some exhaustive disjunction such that 'x or y or ...' implies (any) t, and the latter is reducible to the former?

Let me turn to the part of Cartwright's reply which I find most persuasive. Against DC, Cartwright believes that there is no such exhaustive disjunction. Pushing, pulling, promoting, inhibiting, etc. are kinds of causing. Each can in principle obtain in indefinitely many different ways depending on the context. And if there really are indefinitely many different relations referred to by the label "causal", 'causes' can only be analysed as an *open-ended* disjunction. This, in turn, makes the disjunctive strategy pointless. Even if one could in principle reduce specific instances of causation to either this or that

set of conditions, one would not thereby get closer to having an analysis of 'causes'.[88] Causality is at most analysable as a *cluster concept*, such that for large clusters of relations more or less the same cluster of conditions is satisfied. But, *contra* Longworth, no reductive analysis is possible: causality has an *excess content* with regard to the cluster of conditions. (I will refer to this view as the 'excess content thesis'.)

For Cartwright, however, this is not a problem. For an account of causation to be informative, in spite of the diversity of the thick causal relations, it must be possible to group the t's as causal in some principled way. This does not depend on the possibility to analyse causality "as a (possibly very long) disjunction" (Cartwright, 2010, p. 327), but rather on a loose family resemblance among the t's. To this, one could add that spelling out the *respect* in which the (meaning of the) t's resemble each other would surely contribute to make causation less mysterious, and TC more appealing. However, as I argue in the next section, the details of Cartwright's reply bring to light not just the weakness of determinate pluralism, but also of—the thick-concept variant of—indeterminate pluralism.

6.4 What is a cluster concept?

Let us grant that causation is, in some sense, pluralistic. Still, I think that DC and TC suffer from the same, crucial problem, provided one agrees that the monist's challenge deserves a philosophical answer: Why do different notions, whether determinate or indeterminate, all count as *causal* notions? What makes the relations referred to by the these notions, in spite of their diversity, all *causal* relations?

Consider DC. What's the rationale behind the disjunctive strategy? To be fair to Longworth, DC allows that there be criteria shared by several disjuncts, so overlapping (or clustering) between the various concepts is possible (see Longworth, 2010, p. 314).[89] But even if there is as a matter of fact such an overlapping, whether partial or total, DC *does not explain*—let alone justify—in a principled way the being causal of all those relations by reference to the

[88] Also, as I explain in chapter 7, a particular inference from x to c may be correct, without c being reducible to x. In fact, due to non-monotonicity, the inference may turn incorrect given some appropriate strengthening of the antecedent.

[89] Actually, DC also allows that there be a set of criteria (in the sense of INUS conditions) shared by *all* disjuncts, hence necessary for causation itself. That is: IF $(AX \vee BX \ldots \leftrightarrow C)$, THEN $(C \rightarrow X)$. What Longworth, being a pluralist, does *not* allow is that X be also sufficient for causation.

criteria they have in common.

A similar objection can be levelled against TC. If we believe an account of causality should explain *why* different sorts of causal relations are all causal, TC does not provide a sound alternative to DC. We may, then, be lead to—develop, if it is not on offer—a 'more sophisticated' form of causal monism, viz. inferentialism. First, however, let me explain what is wrong with TC.

Let us consider the way Cartwright motivates her version of indeterminate pluralism. "If there is no universal account of causality to be given, what licences the word 'cause' in a law? The answer (...) is: thick causal concepts" (Cartwright, 2004, p. 806). And what, in turn, grants the family resemblance among the thick causal concepts? Answer: the fact that (the cluster of) thick descriptions often implies (a cluster of) similar features characterising the conditions under which causal laws work (see Cartwright, 2010, p. 327). Which I interpret as follows: the thick causal concepts all imply causation (i.e., are all *causal*) because when they are correctly applied a cluster of conditions is typically satisfied that, as a matter of fact, is also satisfied when the thin concept 'causes' is correctly used. Not much of an *explanation*. First, since for TC each thick verb refers to a different kind of relation, family resemblance may be too weak to explain why the thick verbs form *one* cluster. Secondly, and more importantly, family resemblance among thick concepts does not say why certain concepts are *inside* the cluster and other concepts are *outside*. But perhaps Carwright's reply wasn't meant to explain, only to *describe* the family resemblance. Either way, as I am going to argue, this leaves the indeterminate pluralist's position open to the monist attack.

Let us assume that the correct applicability of thick causal descriptions can be—typically, if not always—decided on the basis of a unique cluster. If the reason why the thick concepts are all causal is that they typically involve the same cluster of criteria, it is not clear why the thick concepts themselves are so indispensable. If all thick concepts typically entail causation plus the same cluster, it would seem that what makes causation special is the cluster itself. So, why not reduce causation to the cluster, bypassing the thick concepts? Now, whilst I do agree with Cartwright that 'causes' as well as thick concepts cannot be analysed in terms of if-and-only-if truth conditions, I think a defense of this conclusion requires an argument not based on TC itself, but on the way a cluster concept is best analysed.

As I said, for Cartwright as well as many other indeterminate pluralists, 'causes' is a cluster concept. I agree on this. But how should a cluster

concept be analysed? I propose that the cluster be analysed in inferential terms. Before explaining *how* this could be done, some clarification is needed as regards the notion of 'cluster concept'. I follow Godfrey-Smith (2009) in drawing an analogy between a cluster concept and a 'jumble' of tools (or criteria, or test conditions). When applying a cluster concept, "[d]ifferent people are free to weight different tests differently, and free to use different weightings on different occasions" (*ibid.*, p. 331). There is a clear sense in which a cluster may both constitute a diverse and a vague concept: it can be *diverse* insofar as its correct application is guided by a jumble of criteria; and it can be *vague* due to the vague applicability of many, if not all, criteria that belong to the jumble. Still, so Godfrey-Smith claims, the cluster can be *one* because, even if the criteria sometimes pull apart, they 'typically' don't—although the typicality itself, arguably, comes in degrees. The above observations parallel more general considerations on the nature of concepts, considerations usually invoked to undermine the 'classical' view that concepts have a definitional structure, explicable in terms of necessary and sufficient conditions (Laurence and Margolis, 1999, §2). On the one hand, the unclear applicability of the criteria in the cluster and the lack of clear intuitions on how to judge instances that satisfy some criteria but not others make it sometimes hard to decide what belongs to a cluster and what doesn't. In this sense, 'causes' is vague, or 'fuzzy'.[90] On the other hand, the judgment that an instance belongs to the cluster comes in degrees of typicality, depending on how many criteria are jointly satisfied. The fact that many relations are such that several criteria are jointly satisfied makes them act as 'prototypes', or conceptual core, with respect to which the belonging to the cluster of borderline relations is evaluated.

I leave to §6.7, after the introduction in §6.5 of the inferentialist framework, the discussion of whether these criteria are best interpreted as evidential conditions, which make the cluster concept *vague*, as Godfrey-Smith suggests, or as distinct concepts, which make the cluster concept *unspecific*, and only its applicability conditions vague, as argued by Reiss. In the remainder of this section, I argue against Cartwright's view that clustering is best interpreted as family resemblance among distinct kinds of physical causation, each kind being more informative than 'causes'.

Are 'thick' concepts more or less informative than, or equally informative

[90] This kind of vagueness should be distinguished from the vagueness of concepts prone to figure in a sorites series due to the unclear applicability of *one* criterion, e.g., observational concepts such as 'bald' or 'red'. More on this in §6.7.1.

as, the 'thin' notion of cause? In what respect and to what extent do they resemble each other? These questions cannot be easily answered. In general, this depends on the *use* the various concepts are put to. In order to evaluate this, we cannot just take into account the conditions under which it is appropriate to infer the causal claim. We must also consider the consequences that ensue from the appropriate application of the claim.[91] Intuitively, we can regard a claim as more or less informative depending on the number and kind of claims which are entailed by it and are incompatible with it. As far as causation is concerned, what we expect from knowledge of causal relations, whether thin or thick, is that they enable correct predictions, explanations and interventions.

Let us assume, as is plausible, that to conclude that a thick fact takes place one needs different evidence from that required to conclude that causing takes place. So, in a given case we may know that there is transference of conserved quantities, probability raising, counterfactual dependence, a mechanism, etc. This typically legitimates the inference to the causal claim. Still, we wouldn't know whether it is pushing, or pulling, or breaking, or binding, or buying, or selling, or... that takes place. Additional knowledge is required. Sometimes this knowledge can be observational—in the case of, e.g., pushing and pulling among objects. More often, it is domain-specific—in the case of, e.g., breaking and binding among biochemical reactants, or trading among economic agents. That is, in most cases we need to know the typical modes of interactions among the entities in a given domain, e.g., that economic agents typically buy and sell whereas biochemical reactants typically break and bind. However, this only establishes that thick concepts can be appropriately used in different circumstances, not that they have more causal content than 'causes'. To the latter end, one must also show that thick concepts allow more correct predictions, explanations and interventions than 'causes'.

Assume we have established the thick claim 'p53 promotes apoptosis', and that this establishes the thin claim 'p53 causes apoptosis', too. To what use can we put the two claims? As I said, knowledge of causal relations

[91] It is important to stress the significance of this move. Drawing attention to *consequences* of appropriate application, too, results in a further shift in the discussion on the meaning of causal claims. The first shift, from the search of truth conditions to the search of test conditions, was made necessary by the observation that monocriterial analyses do not exhaust the meaning of causal claims, but provide at best *evidence* for causal claims. This second shift, instead, consists in enlarging the class of claims which are constitutive of the meaning of causal claims, so as to include claims that *follow* from the correct application of a causal claim. As I explain in §6.5, this amounts to endorsing an *inferentialist* semantics.

should enable correct predictions, explanations and interventions. How does knowledge of thick facts fare with respects to these goals, vis-à-vis knowledge of thin facts? Arguably, if we know that p53 *promotes* apoptosis, we can expect p53 to raise the probability of apoptosis. But this expectation may be as reasonable as the expectation that p53 will *not* raise the probability of apoptosis, provided further background information is provided. For instance, if we know that p53 is mutated, we will expect that the internal pathway to apoptosis won't work properly. Instead, if we (only) know that p53 *causes* apoptosis, we know that p53 has the *disposition* to cause apoptosis, although we cannot be certain on whether this disposition will be manifested. However, if we know more about the context, we may have more reasons to expect one outcome rather than another. In sum, the two claims may be *equally* informative—*depending on what else we know*.

Nor is it too far-fetched to think that in suitable contexts—in the presence of the right premisses—thin talk may be *more* informative than thick talk. In such cases, we rely on *collateral reasons* to apply, or not to apply, the claim to a broader class of circumstances. For instance, we often realise that claims, or models, which work well in certain circumstances, are successfully exportable to other circumstances. The reason why we notice this is that we draw an analogy between the kind of situation in which the claim was initially formulated and the kind of situation to which we want to export it. The analogy may be well founded, based on, e.g, evolutionary grounds (same/different ancestors), geometrical reasons (same/different topologies), or else. If the analogy is plausible, we then proceed by rephrasing the claim and modifying the context description in the terminology appropriate to the target situation. Crucially, such a move is sometimes *facilitated* by the use of thin talk, and wouldn't be regarded as equally plausible if thicker descriptions were used instead. Hence, the use of the less committal and more general 'causes' can prove not only equally informative, but also more informative, if it suggests inferences that thick notions do not suggest.

Let us consider once again the claim 'p53 causes apoptosis'. If this can, in some circumstances, be conducive to more numerous and/or more successful inferences than 'p53 promotes apoptosis', then the TC tenet that thick concepts are always more informative than 'causes' is undermined in complex systems. In general, thick causal verbs seem to have a more intuitive content, but also narrower applicability. The scope of 'pushes' and 'pulls', for example, is usually limited to one-off activities. 'Promotes', 'inhibits', etc. tend to have

larger applicability, since they refer to more complex activities, that may take some time for completion, involve several intermediary steps, etc. The more general and less committal 'causes' may have even larger applicability. In fact, although 'causes' may not be as informative as thick verbs as regards, e.g., the *net effect* of the relation (whether it is positive or negative) or its *strength* (e.g., 'boosts' seems stronger than 'causes'), when used in context 'causes' is flexible enough as to *acquire* such connotations, the description of the context determining whether the causal effect is positive or negative, strong or weak, etc. Since contextual factors are of crucial importance for the coming about of complex phenomena, TC is not so illuminating when applied to disciplines such as system biology or computational economics. Here, the semantics of causal claims is better explicated by reference to test criteria—which grant the correct applicability of the claim—as well as use criteria—which suggest how to apply it once it has been warranted. So, although the cluster-concept view is in principle compatible with Cartwright's metaphysical pluralism, I see no advantage in holding this latter position with respect to the task of explicating the meaning of causal claims in complex systems.

Let me summarise the results of my discussion so far. First, it is desirable to find a way to analyse causation not as an exhaustive disjunction of concepts but rather as a cluster concept. Secondly, we may want to do so by linking the concept of cause to the other concepts it is related to without reducing it to—or identifying it with—them, in other words, by preserving its status of cluster, of object with vague and flexible boundaries. Thirdly, an appeal to the loose family resemblance among the various causings, as characterised by thick causal notions, is not likely to help in characterising the cluster. I have suggested that significant progress can be achieved if we enlarge the set of propositions which are constitutive of the meaning of causal claims so as to account not only for the conditions in which it is appropriate to infer the causal claim but also for the consequences of its appropriate application. This amounts to endorsing INF, viz. the inferentialist approach to causality.

6.5 Causality as inference

I propose that the cluster be analysed in an inferentialist framework, where the question 'How should the cluster concept be analysed?' translates into the other question, 'Under what conditions are inferences to and from claims involving the cluster concept warranted?'

The question 'How should the cluster concept be analysed?' is by no means the only question that we may expect inferentialism to answer. Other, more specific questions arise, such as: Is it possible to order the causal content of the concepts in the cluster? What, if anything, characterises the cluster that applies to causation in complex systems? I discuss the details of my inferentialist proposal in chapters 7 and 8. Here, I limit myself to argue that an inferentialist analysis of causality need not be strongly pluralist.

Inferentialism is an approach to semantics (Harman, 1999; Dummett, 1991; Brandom, 1994b) rooted in the pragmatist tradition (Sellars, 1953; Wittgenstein, 1978). Brandom's own version of inferentialism—to which I'll make reference in the following—explicates the meaning of linguistic expressions in terms of "what can both serve as and stand in need of *reasons*" (Brandom, 2007, p. 654), that is, in terms of *inferential* relations between circumstances of appropriate application (*premisses* of inferences) and appropriate consequences of application (*conclusions* of inferences).

Expressions derive their meaning from the rules of inference they obey (see Brandom, 2007, §1-§3). In particular, subsentential locutions derive their meaning not from their referential function but from the sentences in which they occur[92]; and sentences, in turn, derive their meaning not from their truth conditions but from their inferential role (cf. Sellars, 1962, 1968). The meaning of a sentence (e.g., a specific causal claim such as 'X causes Y') as well as the content of a concept are analysed in terms of meaning-constitutive inferences that, respectively, *warrant* the applicability of the sentence/concept and *are warranted* by it. Following Reiss (2011), I call the sentences that warrant the claim 'inferential base' and those warranted by it 'inferential target'.[93] We needn't be too concerned with the details of the inferentialist semantics for the moment, only bear in mind that the inferentialist has a machinery to get the meaning of words (e.g., 'causes') out of inferences.

Now, if one endorses inferentialism about meaning, what should one say about the meaning of causal claims? Arguably, the meaning of causal claims

[92]The former approach is adopted by 'conceptual atomism', a theory of meaning which extends to all concepts Kripke's account of the meaning of proper names in terms of direct reference; the latter approach, instead, is embraced by so-called 'classical-', 'prototype-' and 'theory-' theories of concepts (Laurence and Margolis, 1999).

[93]In particular, the class of inferences must comprise not just the logically correct ones but also the *materially* correct ones, and not only language-to-language inferences, but also inferences involving non-inferential circumstances of appropriate application (observations) and non-inferential appropriate consequences of application (actions) (see Brandom, 2007, pp. 657-658).

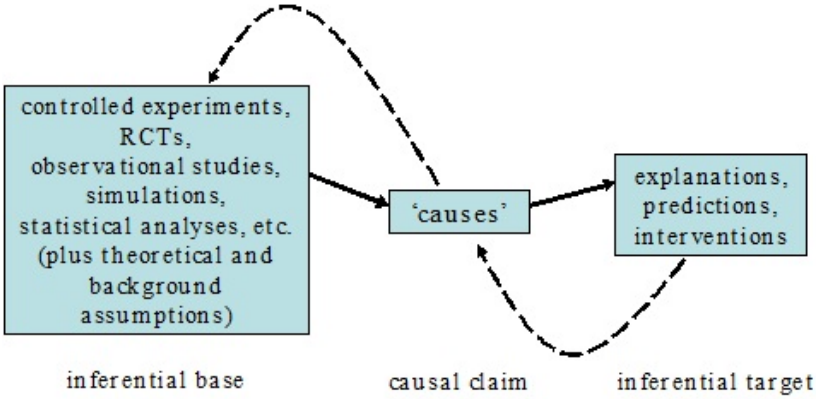

Figure 6.1: Causality *as* inference

is constituted by and analysable in terms of inferential relations between evidence of experiments, results of RCTs, observational studies, simulations, statistical analyses, etc. (together with theoretical and background assumptions) on the one hand, and possible explanations, predictions and interventions on the other (figure 6.1). But is this sufficient to say what 'causes' means?

Problems may arise here (§6.6), having to do with identifying the meaning-constitutive inferences of, respectively, a *particular* causal claim and 'causality' *simpliciter* (i.e., 'causes' as occurring in *all* causal claims). At the root of these problems is a traditional trouble for inferentialism, viz. semantic holism: If the meaning of linguistic expressions is inference, and inferences are all related to one another, then the meaning of *any one* linguistic expression depends on *all* inferences (Quine, 1951). How can one isolate the meaning-constitutive inferences from the non-meaning-constitutive ones?

In the case of causality, the problem may be particularly acute. Intuitively, holism is easier to address in the case of observational statements and expressions, which are 'low-level', viz. closer to the observational contexts in terms of which their meaning can be made explicit.[94] But 'causality' is (often) 'high-level', and related to observational contexts only 'indirectly', through many other inferences. All the more reason to believe that 'causes', whether in a specific claim or *simpliciter*, is so entrenched in our conceptual apparatus that its meaning-constitutive inferences are too 'distributed' to be isolated.

Now, *whether* the inferentialist approach is *in general* successful depends on whether a *general* response to holism is successful. The inferentialist may

[94]Yet, it seems possible to define inferentially *logical* vocabulary (Dummett, 1991).

claim that meaning is fixed by the counterfactually robust inferences (Sellars, 1948), or by canonical introduction and elimination rules, and the possibility to justify other inference rules by reduction to such rules (Dummett, 1991), or that communication is possible in virtue of similarities rather than identities of inferential role (Harman, 1999), or the distinction between one's commitments and the commitments attributed to others (Brandom, 1994b), etc. Depending on one's favourite take, there's either a sharp distinction between meaning-constitutive and non-meaning-constitutive inferences (non-holists, such as Sellars and Dummett) or a blurry one (holists, such as Harman and Brandom). Which approach one should choose depends, among other things, on whether one can make use of the distinction between confirmational and semantic holism (Cozzo, 2002). Here I don't commit to any of the stronger replies, I only limit myself to observe that there are only few examples of allegedly successful reductive analyses, and the intuitions such analyses rely on are seldom shared among all philosophers (Margolis and Laurence, 2012, §2.1, §5.2). The failure of monistic and determinate pluralist analyses of causality suggests that sharp distinctions may not be available for *all* concepts. Accordingly, in the following I will only assume that some weak response is available such that semantic holism may be 'contained', at least for a class of concepts, or for some concepts better than others. On this assumption, we still need an *additional* argument to the point that causality is among those concepts for which an inferentialist analysis is informative. Providing such an argument is apparently harder because of the centrality of the notion of causality in our conceptual apparatus. Still, I want to argue that there is room for a weak form of semantic monism.

I will now present in some more detail the evidential and the semantic versions of pluralism and their connections with inferentialism, before discussing whether and to what extent they follow from the inferentialist approach to meaning (§6.7).

6.6 Evidential vs semantic pluralism

Should we endorse only *evidential* pluralism (Williamson, 2006; Russo and Williamson, 2007) or also *semantic* pluralism (Reiss, 2009a, 2011)? For the evidential pluralist, traditional monistic accounts offer at most evidence for causal relations, i.e., test conditions, *not* exhaustive analyses, i.e., truth conditions. Provided one rejects such accounts, this claim is quite straightforward

and uncontentious. More contentious is whether these different analyses identify distinct notions of cause. The debate has an obvious bearing on whether the attempt to provide a monistic account of causality is well-founded.

On the one hand, in line with his epistemic view, Williamson claims that different kinds of evidence for causal claims do not presuppose different concepts of causality. The evidence gathered by using the various criteria, which 'typically' apply together, is evaluated depending on how well it conduces to successful uses—viz. successful predictions, interventions and explanations—so as to *univocally* constrain rational belief in what causes what.

On the other hand, Reiss suggests the following identity condition for the meaning of 'causes':

> Suppose the term 'cause' is used on two different occasions and it is not known whether it has the same meaning on both occasions. Two such claims would have the form 'X α-causes Y' and 'Z β-causes W'. We can then say that 'α-causes' has the same meaning as 'β-causes' (on these occasions) to the extent that 'X α-causes Y' is inferentially connected to the same kinds of propositions regarding the relation between X and Y as 'Z β-causes W' is inferentially connected to propositions regarding the relation between Z and W (Reiss, 2011, pp. 923-924).

This condition relativises the meaning of a token causal claim to both *kind* and *number* of criteria the claim is inferentially related to—where 'criteria' stands here for both *test* conditions (base criteria) and *use* conditions, or purposes (target criteria). One thing, in fact, is to say what 'causes' *simpliciter* means, quite another thing to say what 'X causes Y' means when applied to population P_1 as opposed to population P_2, since, e.g., mechanisms or probabilities by which the effect is brought about as well as criteria used to establish the claim may be different in P_1 and P_2.

For this reason, the meaning of any causal claim—e.g., in epidemiology, 'exposure to aflatoxin causes liver cancer'—is population-relative:

> in general, when the causal claim concerns the toxicity of a substance, language users are entitled to inferences about a given population only when the inferential base contains evidence claims about just that population (Reiss, 2011, p. 917).

To this consideration, one may add that identity depends not only on sameness of *kinds* of sentences, or (formal) criteria, but also on sameness of *contents* of sentences, or 'propositions'. This further condition seems implied by what Reiss himself elsewhere states: "[t]here is a definite set of propositions with which any causal claim is inferentially related" (Reiss, 2011, p. 924).[95] Notice that this applies even when the same relata, hence the same causal claim type (as in, e.g., 'X α-causes Y' and 'X β-causes Y'), are concerned. That is, meaning depends not just on base and target criteria but also on the specific content of base and target sentences. Indeed, the content of the sentences need not coincide with the content of the criteria. For instance, one thing is to say how a randomised control trial (RCT) is inferentially related to a causal claim, i.e., how it is to be conducted to provide support for the claim. Another thing is to specify how the meaning of 'exposure to aflatoxin causes liver cancer'—whether applied to *different* populations (e.g., mice and humans) or to the very *same* population—depends on the particular way the RCT is conducted, i.e., the specific individuals on which the RCT is performed (each, arguably, instantiating slightly different mechanisms), the result of the randomisation in a specific trial, etc. So, even on the assumption that the criteria are the same, meanings may end up different depending on the different circumstances in which tests are performed and on the different actions taken—described, respectively, in base and target.

Now, since meaning of both criteria and contents is to be defined inferentially, it is easy for pairs of inferential webs to differ somewhere. No surprise that Reiss takes inferentialism to entail conceptual pluralism: the concept of cause is "*unspecific* rather than *ambiguous*" (Reiss, 2011, p. 914), and should be replaced by a plurality of well-defined—or well-defin*able*—concepts.

Who is right? In spite of the disagreement, both Williamson and Reiss maintain that causality has to do with inference. Shouldn't we expect from inferentialism itself to deliver an answer on the monism-vs-pluralism debate? Well, things are not so easy. Two readings, in fact, are available.

According to a first reading, clearly adopted by Reiss, 'causes' is ambiguous and differs from claim to claim, depending on the particular base and target. That is, tokens of 'causes' used in claims established by different methods and licensing different inferences have different meanings.

[95] Admittedly, the *content* of base and target sentences may not count, for Reiss, as part of the identity conditions for 'causes'. Indeed, that he intends to impose this further condition is denied by him (private communication). Still, *pace* Reiss, his argument for the unspecificity of 'causes' *presupposes* this condition (§6.7.1).

A second reading, instead, has it that the premisses from which causal claims are entailed, and the conclusions that causal claims entail, only differ as to the *weight* of the different kinds of inferences which, respectively, warrant the claims and the claims warrant. Causality is only 'moderately' diverse: criteria employed in different circumstances, although weighted differently, are similar enough to legitimate a strong clustering among the various tokens of 'causes'. This second reading, for which causality is *one* 'cluster concept', is closer to Williamson's position.

The issue is all the more relevant insofar as, according to one's favourite reading of the inferentialist take on meaning, evidential and semantic pluralism give two different answers to the question of what, if at all, is the philosophically interesting fact that explains the use of "causal" to denote all kinds of causal relations. As mentioned, for Reiss there is no fact of the matter: whether different criteria 'typically' coincide is an empirical not conceptual matter, "much like discovering that various symptoms of a disease typically co-occur" (Reiss, 2009a, p. 33) (cf. Psillos, 2010); "why we have come to call the different kinds of relationships causal is a matter of historical, not philosophical, inquiry" (Reiss, 2009a, p. 37). For Williamson, instead, there is *one* fact of the matter: different criteria typically coincide because they all provide evidence for the *same* (kind of) relation; we have come to call all these relations 'causal' because they *all* share the role of licensing inferences. Which answer is correct?

The issue largely depends on whether 'causes' is to be regarded as *vague*, as the monist can maintain, or *unspecific*, as Reiss claims. If 'causes' is vague, there can be *one* cluster. In this case, vagueness is *semantic*. If 'causes' is unspecific, instead, there are irreducibly *many* notions. On this view, the 'vagueness' is *epistemic*, i.e., it concerns ignorance on which of many distinct notions is employed in each case: "True, we might not always have a very clear idea of what [the meaning-constitutive sets of propositions] are. But this is a question of epistemology, not of semantics" (Reiss, 2011, p. 924).

One way to tackle the issue is to look at the foundations of the inferentialist project itself. As mentioned, even granting that holism can be contained, we still need an argument to show that 'causality' is liable to be analysed inferentially.

More specifically, holism generates two further problems: (i) the *(in)stability of conceptual contents* under change of belief and commitment to the properties of various inferences; and (ii) the *(im)possibility of communication*

between individuals who endorse different claims and inferences. But, then, how can a change in one belief not result in a change in all other beliefs? And how can two speakers' whose beliefs differ somewhere ever talk about the same thing?

When it comes to causality, the inferentialist needs an argument to guarantee, in the face of the holistic nature of meaning, (i) the relative stability of conceptual content of 'causes' and (ii) the possibility of successful communication on causal claims. In §6.7, I show that, if these two goals can be achieved, holism can be contained in a way that makes room for 'causes' being *one* and *vague* rather than unspecific. I offer two arguments to this point, one for the vagueness of the notion of cause—the 'argument from (in-)stability of content'—another for its uniqueness—the 'argument from communication'.

6.7 Causality *as* one, vague concept

6.7.1 Argument from (in-)stability of content

To begin with, notice that Reiss' argument for semantic pluralism trades on an ambiguity between (i) analysing meaning in terms of different *sets* of (token) sentences and (ii) analysing it in terms of sets of different *kinds* (or types) of sentences within those sets.

On the first interpretation, strong semantic pluralism follows straightforwardly. To each inferential base and target of (token) sentences there correspond a different meaning. However, this would make it strictly speaking impossible for linguistic expressions to share the *same kind* of meaning—unless some similarity criterion is allowed, that is, unless one opts for the second interpretation. Intuitively, it must be possible and legitimate to group linguistic expressions depending on their similar inferential roles. Analogy of inferential roles could then be used to make meaning of kinds of linguistic expressions explicit. Otherwise one would be stuck with (meaning of) single-case claims and unable to formulate general claims.

The second interpretation is surely more attractive. Here, strong semantic pluralism follows only on the assumption of a 'double standard' for the semantics of 'causes' and the ancillary notions (e.g., 'depends', 'produces') that help make 'causes' explicit. That is, if one wants to distinguish between *kinds* of sentences—which is plausible—and still draw the same pluralist conclusion on the meaning of causal claims, one must have a similarity criterion

that legitimates the clustering of sentences in base and target of the ancillary notions but prohibits the clustering of sentences in base and target of 'causes'. For instance, the criterion will dictate that the circumstances that entail and are entailed by 'probabilistically depends' are similar enough to make *one* cluster, whereas those that entail and are entailed by 'causes' are not. As a result, 'probabilistically depends' counts as one concept, whereas 'causes' is an unspecific term that subsumes a plurality of specific concepts. But this is problematic, as I am now going to show.

For the inferentialist, language is essentially *dynamic*. This means that

> any codification or theoretical systematization of the uses of (...) vocabulary-kinds by associating with them meanings that determine which uses are correct will, if at all successful, be successful only contingently, locally, and temporarily (Brandom, 2008b, p. 5).

Drawing on this insight, the thesis of the argument from (in-)stability of content (*Sta*) is that the respect in which different meanings differ cannot be made fully explicit not only for epistemic but also for semantic reasons:

> ***Sta*.1**. Communication can be unsuccessful not only because of the ignorance of the speakers but also because language is a *dynamic* network of concepts.
> ***Sta*.2**. If so, then the meaning of 'x-causes' can be specified only *on-the-fly* by successful use.
> ***Sta*.3**. But, at any time, *also* the meanings of the ancillary notions that should make 'x-causes' semantically explicit is specified on-the-fly.
> ***Sta*.C**. Hence, the meaning of 'x-causes' can be—semantically—only *partially* specified, or specifi*able*. That is, it is not only semantically unspecific but also *semantically vague*.

To claim otherwise would mean to accept the double standard: meaning of ancillary notions is vague whereas meaning of 'causes' is unspecific. But how could this be justified?

True, the meaning of ancillary notions may be easier to make explicit. In inferentialist terms, their tokens are easier to group as belonging to the *same kind* on the basis of their inferential role. For instance, the meaning of 'probability raising' or 'counterfactual dependence' can be (following Brandom)formally fixed in terms of necessary (target) and sufficient (base) conditions. However, holism applies to such concepts, too: when it comes

to applying the formally-defined concept to *non-formal* circumstances, it is still left to us to decide whether the formal notion applies or not—whether 'X raises the probability of Y' or 'Y counterfactually depends on X', etc. Strictly speaking, *no* concept can be *totally* isolated from the other concepts.

And if one applies to the ancillary notions the same strict identity condition which is applied to 'causes', one ought to conclude that distinct tokens of the ancillary notions (e.g., 'α-probabilistically raises' as occurring in 'X α-probabilistically raises Y' as opposed to 'β-probabilistically raises' as occurring in 'Z β-probabilistically raises W') have distinct meanings, since their inferential role is fixed by distinct sets of sentences. Token sentences in base and target are different for *each* claim—whether involving the concept of cause or other concepts.

So Reiss' argument for unspecificity generalises in principle to the ancillary notions: one cannot say there are many distinct, unspecified concepts of causality without at the same time saying that there are many distinct, unspecified concepts of probability raising, counterfactual dependence, etc. From which it follows that if we want the meaning of the ancillary notions to be semantically vague on *similarity* grounds, then we should allow causality to be vague as well, on the same grounds.

Admittedly, one could object to this analysis that the vagueness of 'causes' is of a different kind from that of the ancillary notions, 'causes' being more like a multiply-realisable concept (e.g., 'bird', realised by distinct species of bird), the ancillary notions more like *non*-multiply-realisable concepts (e.g., 'bald'). But even granting that this is so, we don't deny that there is one, legitimate concept of 'bird-ness'.[96] So why deny that there is one concept of 'causality'?

Now, a serious problem may arise, having to do with the possibility of meaningful communication. If the identity of meaning depends on the inferences in terms of which the meaning is analysed, and is only partially fixed/fixable, how can meaning be stable enough to allow us to discuss about the same things? In particular, how can meaning be stable enough so that we can make it explicit in communication, by clarifying the inferential presuppositions and implications of what we say, so as to resolve controversies on what claims to endorse?

[96] To reiterate a point made earlier (see fn. 102), I am not concerned here with metaphysical issues, such as whether or not there is an object-type the concept refers to, whether bird-ness is a 'natural' kind, whether causal relations are 'real', etc.

Clearly, due to holism, conceptual/propositional content cannot be *totally* transparent to the speakers (see above). Yet, much of what we say, in particular many of the implications of our commitments, *must* be transparent to us—otherwise communication wouldn't be possible. Granted that holism can be contained, there must be a 'semantic level' at which meaning of many notions that we use in communication is stable enough. Arguably, that is the level at which meaning is *vague enough* for the speakers to take it as *one* and treat its possible applications as *similar enough*. That is the level at which, I claim, there is a *unique* concept of causality. But here we need an argument to the point that we can legitimately talk of *one* vague notion of causality, as opposed to *many* vague notions.

6.7.2 Argument from communication

Before presenting the argument, it is crucial to point out the source of the difference between Reiss and Williamson in the unity-vs-disunity debate, namely the different weight they ascribe to base and target.

On the one hand, Williamson places more weight on the target. All kinds of evidence are required, in principle, to establish a causal claim (Russo and Williamson, 2007). This is because, ultimately, their role is helping maximise the target's success. The meaning-constitutive inferences are (only) claim-to-target inferences (see Williamson, 2006, p. 78). And since the target is a unique class of claims, namely explanations, predictions and claims about results of interventions, there is *one* notion of cause. That is, the unity of purpose(s) is what fixes the meaning of 'causes' and blocks the fragmentation. Here, inferentialism is used to explain philosophically the uniqueness of meaning.

On the other hand, Reiss places more weight on the base: "whereas the meaning of an expression is given by its inferential connections (...), its method of verification determines what these inferential connections are" (Reiss, 2011, p. 923). And since we have many ways to establish a causal claim, i.e., many appropriate base-to-claim inferences, we have *many* concepts of cause as well (see Reiss, 2011, p. 924). Nor does an appeal to purposes (target) help unify. Rather, purposes *dis*unify, by pulling apart the role of causal claims. So, an epidemiologist may be interested in explaining whether the population-level correlation between aflatoxin exposure and liver cancer is due to the carcinogenicity of aflatoxin; a policy maker in knowing whether

controlling aflatoxin is an effective way to reduce mortality; someone exposed to aflatoxin in predicting whether this exposure will result in a higher chance of liver cancer. To the three purposes there correspond different criteria to establish the causal claim, hence different concepts of cause (see Reiss, 2011, §3-§5)—even within the very same discipline.[97] The fact that different relations are all called 'causal' has no philosophical significance. Inferentialism is only invoked to describe the fragmentation, not to explain the way this came about, which is a contingent, historical matter.

Here, it is worth contrasting Reiss' position with Godfrey-Smith (2009)'s. For Godfrey-Smith, a cluster concept is an "amiable jumble" of criteria, such that its criteria sometimes pull apart but 'typically' don't. Causality is peculiar because the jumble is "cantankerous", not amiable (Godfrey-Smith, 2009, p. 331). By this, he means that causality is partly an "essentially contested concept" (ECC) and partly low-level, or uncontentious. It is ECC, because

> it is not just hard to work out when the conditions of application are met, but (...) the conditions for application themselves are, given the concept's role, permanently susceptible to being challenged and renegotiated (Godfrey-Smith, 2009, p. 335)

However, it has also low-level, uncontentious uses, where disputes with respect to boundaries and criteria for application *can* be resolved. Terms with both ECC and low-level uses acquire their role when

> their successful application has significant downstream consequences, but their domain is complex in ways that involve the absence of sharp boundaries that function as attractors to usage. [In the case of causality], an accepted set of exemplars and a sense of shared *purpose* behind diverse uses prevent a fragmentation into distinct concepts. These ideas might be linked to tools developed in recent 'inferentialist' philosophy of language (...) (Godfrey-Smith, 2009, p. 336)

In virtue of such a shared sense of purpose, unity prevails over disunity. This

[97] Reiss (2009a, §6) articulates this idea as applied to economics: (i) when the purpose is prediction, causal talk usually refers to a property of time series such that one is a good predictor of the other ('Granger causality', see (Granger, 1969)); (ii) in policy claims, causality is best interpreted as stability of the relation between two variables under intervention (Haavelmo, 1943); (iii) causality as used for explanation has mostly a mechanistic sense (Elster, 1998).

conclusion, partly reached by invoking inferentialism, is in obvious disagreement with Reiss, for whom an appeal to purpose does *not* prevent fragmentation.

How should one decide between evidential and semantic pluralism? To answer this question, partly drawing on Godfrey-Smith's distinction between ECC and low-level uses, I employ the argument from communication (*Com*):

> ***Com.*1**. For distinct concepts of causality to exist there have to be different communities using the *same* word with *different* meanings, and yet either (i) *never* successfully communicating by using the word, viz. never agreeing whether implicitly or explicitly on its rules of 'correct' application; or (ii) communicating by using the word but *never* discovering the disagreement on such rules.
> ***Com.*2**. The second option is clearly implausible, as evidenced by the explicit, semantic disagreements among both scientists and philosophers.
> ***Com.*3**. The first option is implausible, too: true, often different communities don't discuss subject-specific causal claims; but sometimes they do; plus, sometimes they engage in high-level semantic reflections; in both cases, they *can* reach semantic agreement.
> ***Com.*C**. Therefore, 'causes' *can* be—in a sense to be qualified—*one* concept.

This argument agrees with Williamson that—in a sense—purposes contribute to unify, whilst conceding to Reiss that—in another sense—they don't. What exactly are the two senses can be illustrated in terms of different *semantic levels*, i.e., different contexts in which purposes play their role in fixing the meaning (see below). Depending on the level, we may have either unity (uncontentious uses) or disunity (contentious uses). In both low- and high-level cases, a shared purpose—together with the taking of our linguistic expressions as *committing* us to certain consequences[98]—tends to generate a shared commitment on meaning. That is, several speakers/communities that given some accepted base agree on endorsing a claim *ought to* agree on the claims that follow from it and the claims that are incompatible with it—that is, they ought not deny (respectively, endorse) the claims that follow from it (are incompatible with it), once they are made aware of them.

First, I agree with Godfrey-Smith that unity is possible at the subject-specific, low level of *token*, 'C causes E' claims. This may not be the case when, say, distinct claims (e.g., 'X causes Y' and 'Z causes W') are concerned. Here, different purposes can easily produce low-level variability. However,

[98] This is usually referred to by the inferentialist as the 'normativity of meaning' (Peregrin, 2012).

take the *same* claim, e.g. 'exposure to aflatoxin causes liver cancer', and *distinct* communities holding some commitment (not necessarily the same) towards the claim, e.g., community A endorsing the claim in the light of an RCT in mice, community B not endorsing it, in the light of an observational study in humans. Here, disunity is possible only provided findings of one community about the claim, obtained by using one criterion and ensuing in certain consequences, once communicated to the other community do *not at all* affect their commitment to the claim. But this sounds implausible. For instance, it is implausible that A's evidence for the causal connection between aflatoxin and liver cancer in mice has *no* bearing on B's belief that aflatoxin causes liver cancer in humans. Reasoning from analogy, especially on evolutionary grounds, is common in science. And the possibility of an RCT in mice and not in humans doesn't seem to be reason enough for the meaning of 'causes' in the two claims to be of different *kinds*.

On the contrary, whenever scientists coming from different backgrounds interact, for instance in interdisciplinary projects, they must agree on the interpretation of their results as well as accept the bearing of each others' methods on such results. The outcome of their research is very often the formulation of causal claims. When urged by philosophers to say what they mean by 'C causes E', scientists may well disagree—this largely depends on how their training shapes their methods and purposes. However, they also need to come up with a coherent story. After all, the causal claim is the result of a collaborative effort, shared methodology, assumptions, results' interpretation, etc. There must be something they all mean by 'C causes E', at least in that context.

One instance of such interdisciplinary projects is EnviroGenoMarkers. This project is driven by the success of past epidemiological studies in identifying the risk of environmental exposures (e.g., air pollution and passive smoking) with respect to the onset of chronic diseases (e.g., cancer and coronary artery disease). The current project, whose methodology comprises both experiments on biological samples and observational studies, aims at "the identification of both biomarkers of exposure (e.g. dietary components, environmental pollutants) and of markers of early damage (e.g. early disease-specific metabolic changes), notably for carcinogenesis" (Chadeau-Hyam et al., 2011, p. 84). The team includes chemists, molecular biologists, epidemiologists, statisticians, etc. Due to such a diverse composition, one would expect a clash of intuitions on what counts as causal. In spite of this, there is agree-

ment in interpreting previous studies on environmental exposures as providing evidence for causal claims: "biomarkers (...) have contributed to make the association more plausibly causal" (Vineis et al., 2009, p. 54). And with regard to their own study, the team maintains that "the finding that preclinical biomarkers shown to be related to particular exposures in prospective studies are also elevated in certain subclasses of disease would strengthen causal links between exposures and disease" (Chadeau-Hyam et al., 2011, p. 85). At the same time, the connection between purposes is implicitly acknowledged: "intermediate biomarkers can provide important mechanistic insight into the pathogenesis of environmental diseases" (Vineis et al., 2009, p. 54). That is, a successful prediction provides at least some explanation. This supports my unity-over-disunity claim: to the extent that there is some connection between purposes, successful communication is conducive to low-level unity, not fragmentation.

Secondly—here I part with Godfrey-Smith—unity may be possible also at the non-subject-specific, high level where speakers discuss the meaning of a word, e.g., 'causes' (now treated as a *type*), and agree on the formal criteria that are 'typically' associated with it. In science, such 'abstract' discussions may not take place frequently, but they do take place sometimes. Instead, in areas of discourse such as philosophy they take place on a regular basis. Now, *whenever* they take place between any two speakers A and B, provided the (counter-)examples used by A, who adopts certain rules for the use of some expression, are regarded by B as informative on the correctness of B's rules of application for that expression, the possibility of some unique concept—that is, a shared core of rules—that A and B are referring to is presupposed.

This largely depends on the existence of a stable class of interrelated purposes. In particular, the class must be *stable enough*, so that discussions on the appropriate tools to achieve the purposes, in the light of past consequences of the tools' application, can take place. And the purposes must be *interrelated enough*, so that the success with regard to one purpose, driven by the application of one tool, affects to some extent the other purposes as well. These two conditions seem both well satisfied in the case of causality. Here the class of purposes, i.e., prediction, intervention and explanation, is traditionally very stable. And, although meaning is in principle flexible, such a class seems—although this is just my best guess on the way our linguistic practice will evolve—very unlikely to change. Also, the purposes are strongly interrelated. So, it is hardly the case that a good explanation (say, a mech-

anistic explanation of how aflatoxin causes liver cancer) has *no* bearing on successful prediction, when the appropriate circumstances are in place (when the individual instantiates the mechanism); and it is unlikely that a successful prediction (a preclinical biomarker predicting, in the presence of some exposure, the onset of the disease) tells us *nothing* on possible interventions (either directly, if the biomarker is itself a cause, or indirectly, if disease and biomarker are effects of a common cause); etc. If the above reasoning is correct, then making explicit the meaning of this unique, high-level concept is in principle possible. Causality need not be *essentially* contested. Unity *can* be achieved and, although it is ultimately dependent on variations in tools and purposes, is likely to remain stable.

Conclusion

I examined the various forms of causal pluralism, and found them unable to satisfactorily answer the monist challenge, viz. to explain what the different relations have in common, and why we use only *one* word to denote them. Inferentialism, in contrast, seems to have the resources to face the challenge. The same label "causal" is applied to many, seemingly-different relations in virtue of their sharing the feature of licensing inferences about predictions, interventions and explanations. Endorsing inferentialism need not lead to a strong pluralism on the notion of cause. Although what inferences are licensed in each case may depend on the context as well as the purpose of the enquirer, 'causes' can still have *one*—although vague—meaning across contexts and purposes: as far as our tools and purposes are stable enough (so that we can make them explicit) and related to one another (so that there is something to make explicit), there is, as a matter of fact, one, vague cluster of criteria that helps us best achieve those purposes. Inferentialism encourages unification—*not* fragmentation—of meaning, and helps us understand *to what extent* we are using the same concept, e.g., to what extent the endorsement of 'C causes E' in one context should carry over to another context. In *this* sense, inferentialism can support a 'weak' form of conceptual monism.

Chapter 7

The Inferentialist Account of Causality

In this chapter, I will be concerned with developing an inferentialist account of causality. The leading idea of the chapter is that the meaning of 'causes' is to be interpreted not in terms of the contribution to the truth conditions of the sentences where 'causes' appears, but in terms of the contribution to the correctness of the arguments where 'causes' is part of the premisses or the conclusion. More simply put, the meaning of 'causes' has to do with the inferences it licenses. Accordingly, the account I propose answers the question 'What does 'causes' mean?' by answering 'What inferences are licensed by causal claims?' After presenting the inferentialist framework (§7.1), I address this question by making explicit the meaning of 'A causes B' claims—where the meaning of 'causes' is relativised to specific relata—as well as the meaning of 'causes' *simpliciter*—as this appears in all causal claims, irrespective of the relata (§7.2). Once the general features of the concept of causality have been spelled out, I will be in the position to address the more specific question 'What is special about (the meaning of) causal claims in complex systems?' (see chapter 8). I conclude by discussing how the inferentialist account allows one to deflate the issue of identifying the 'secret' connection underpinning causal relations (§7.3).

7.1 Preliminaries

7.1.1 Incompatibility semantics

One finds the idea that 'causes' is an inference license already in Sellars:

> I shall be interpreting our judgement to the effect that A causally necessitates B as the expression of a rule governing our use of the terms 'A' and 'B' (Sellars, 1949, fn. 2, p. 136).

According to Sellars, 'A causes B' is the expression of a rule. For the inferentialist, modal expressions ('causes' included) have the *expressive* role of inference licenses, and inference licenses are a sort of *rules* (Brandom, 2000, p. 76). Now, why 'expressive', and why 'rules'? Because inference licenses allow us to express in the object language beliefs about relations obtaining in the 'natural space', but are themselves to be analysed in the 'space of reasons', that is, the language of linguistic norms and rules (deVries, 2010, p. 399). For instance, 'A causes B' is an object-language statement that describes the relation between A's and B's. The analysis of the meaning of 'causes' in the statement requires the resources of another language to talk about the object language. And since 'causes' is a 'rule', the analysis must be in *normative* not just descriptive terms. In Brandom's jargon, the analysis requires the vocabulary of 'commitments' and 'entitlements'.

In order to be more precise on how 'causes' should be analysed in inferentialist terms, some qualifications are needed as regards the particular nature of the metalanguage and the relation between metalanguage and object-language in the inferentialist semantics. I will from here onwards mainly refer to Brandom's own way of developing Sellars' insights, which Brandom calls "analytic pragmatism" (Brandom, 2008a,b). Analytic pragmatism unites pragmatism ('meaning is use') with (a kind of) formal, modal semantics. The novelty resides in that semantic analysis is in terms not only of direct relations between vocabularies (e.g., definability, translatability, reducibility, supervenience), but also of 'pragmatically mediated' relations, i.e., relations between vocabularies mediated by linguistic practice, so that relations to linguistic practice become part of the analysis itself.

A paradigmatic example of a pragmatically-mediated relation is the relation of being a *pragmatic metavocabulary*, which holds between one vocabulary (the metavocabulary) and another (the object-language vocabulary) in virtue of some set of practices. More precisely, this relation allows one to say in the metavocabulary what one must do in order to count as saying the things expressed by the object-language vocabulary (Brandom, 2008b, p. 8). For instance, the normative vocabulary of commitments and entitlements acts as a pragmatic metavocabulary with respect to the modal vocabulary, by allowing one to specify how to use terms such as 'necessity', 'possibility', 'causality', etc. Example: the modal necessity in 'Donkeys are necessarily mammals' can be analysed in terms of sentences of the form 'If I were committed to 'My first pet was a donkey', I would not be entitled to 'My first pet wasn't a mam-

mal' '. Most fundamentally, for Brandom the vocabulary of commitments and entitlements allows one to specify how to use 'incompatibility'; in turn, the vocabulary of incompatibilities is a pragmatic metavocabulary that allows one to specify how to use the incompatibility-entailment relation, which becomes the basic relation in terms of which to provide *any* other semantic analysis, whether of logical or of non-logical concepts.[99] In particular, since 'causality' belongs to the modal vocabulary, it must be analysed in terms of incompatibility entailments.

Incompatibility entailments are a sort of counterfactual-supporting, modally robust inferences. The idea is that commitments and entitlements can produce incompatibilities: to say that if one *were to be* committed to p, one *would not be* entitled to q amounts to saying that p and q are incompatible. Brandom's key idea is that semantics is to be based not on the notion of truth, but on the notion of incompatibility—which is why Brandom's semantics is called "incompatibility semantics" (in short, IS). That is, the meaning of a linguistic expression has to do with what ought to be excluded by the appropriate use of the expression, rather than with some alleged correspondence between the expression and certain mind-independent facts. This move is supposed to do justice to the normative dimension of meaning, as something that belongs to the space of reasons, not the natural space.

The inference relation underwritten by incompatibilities is the following:

Incompatibility entailment: *p incompatibility-entails q* iff everything incompatible with q is incompatible with p.

This can be informally understood as saying that *p incompatibility-entails q* iff were one to be committed to p, one would not be entitled to deny q.

Following Brandom (2008a, chap. 5), incompatibility entailment may be formally defined after defining an incoherence property Inc over the sentences of a language $\mathcal{L}$, and an incompatibility function I over sets of sentences. Inc is a subset of the collection of finite subsets in $\mathcal{L}$ that is upward closed, i.e. if $X \subseteq Y$, then if $X \in Inc$ then $Y \in Inc$. Inc generalises inconsistency to the case of non-logical properties, so that if a set is incoherent it remains

[99] Notice that it is Brandom's strategy to prove that it is possible to use the incompatibility vocabulary as the most basic one, and to develop one formal version of analytic pragmatism based on incompatibility, viz. the 'incompatibility semantics' (see below). In the following, I assume that Brandom's strategy is well justified and use his semantics. This involves no commitment on my part on whether this is the only way or the best way of developing the programme of analytic pragmatism.

incoherent when further sentences are added to it. I between sets X and Y, viz. $X \in I(Y)$, is defined as that function such that the union of the two sets is incoherent, viz. $X \cup Y \in Inc$. Now, let us indicate the incompatibility-entailment relation with "$\models_{\text{Inc}}$", and the incompatibility set of a sentence p with "$I(\{p\})$". Then, incompatibility entailment is defined as:

$$p \models_{\text{Inc}} q \text{ iff } I(\{q\}) \subseteq I(\{p\})^{100} \tag{7.1}$$

Since Inc applies also to non-logical properties, and I is an incompatibility of a non-logical nature, the above definition must be understood as a way to cash out the Sellarsian notion of 'material' inferences (Brandom, 2007, p. 657). Material inferences are inferences of the form 'F(a). Therefore P(a)', e.g. 'Lightening now. Therefore, thunder shortly' (Sellars, 1953, p. 323). They are non-enthymematic, i.e., their validity does not rely on the possibility to make the argument deductive by adding implicit premisses, e.g. 'All lightenings are followed by thunders'. In other words, they are valid not in virtue of their logical form but in virtue of the *content* of premisses and conclusions. Also, they are *non-monotonic* (Brandom, 2008a, p. 106): p may commit one to q, but $p\&r$ may not (more on this in §7.1.2).[101]

The propositional content expressed by a sentence p (i.e., its semantic interpretant) may be represented by the set of sentences q that express propositions materially incompatible with p, viz. its *incompatibility set*:

$$\text{Cont}(p) = \{q | \{q\} \in I(\{p\})\} \tag{7.2}$$

[100] Since IS takes 'incompatibility' not 'truth' as basic, in the above definition 'inference' (or 'entailment') is interpreted not as 'truth-preservation' but as 'compatibility-preservation'. In this way, the inferentialist reverses the traditional order of analysis, from explicating 'inference' as 'truth-preservation' to explicating 'truth' as 'what is preserved by inference' (see Brandom (2000, p. 161) and Peregrin (2008, §3)). Also notice that 'inference' and 'incompatibility' are *interdefinable*, although whether 'inference' or 'incompatibility' is chosen as more basic varies from semantics to semantics (Peregrin, 2008, §4). Brandom (2008a)'s own choice is to take incompatibility as basic.

[101] One may wonder why $\models_{\text{Inc}}$ is non-monotonic, i.e. it is possible to turn a correct inference into an incorrect one, but Inc is upward-closed, i.e. it isn't possible to turn an incoherent set into a coherent one. A possible counterexample are abductive inferences, which are apparently such that $p_1, \ldots, p_n \not\models_{\text{Inc}} c$ but $p_1, \ldots, p_{n+1} \models_{\text{Inc}} c$. So, $\{p_1, \ldots, p_n, c\} \in Inc$ but $\{p_1, \ldots, p_{n+1}, c\} \notin Inc$. This is particularly relevant to the case of causal claims, where one criterion alone often does not grant the inference, but several criteria together do. A possible way to account for this is by thinking of abductive inferences as always involving some replacement—and not just an addition—of premisses, viz.: $p_1, \ldots, p_k, \ldots, p_n \not\models_{\text{Inc}} c$ but $p_1, \ldots, p_k^*, \ldots, p_{n+1} \models_{\text{Inc}} c$. Intuitively, the inference becomes correct only after making the context description more precise.

The rationale behind the above characterisation is not to *reduce* the content of a sentence to some fixed set of other sentences with which the sentence is in some special relation, but rather to take it as *useful* to model the content as being so reducible.[102] So, an inferentialist analysis is to be interpreted as a sort of conceptual *explication*, rather than a conceptual reduction. Accordingly, the equality sign signals that whatever is on the right-hand side of the analysis conceptually explicates, not reduces, what is on the left-hand side.

Since the incompatibility set of a sentence is fixed by the incompatibility set of *other* sentences, and so on and so forth, IS is *holistic* (Brandom, 2008a, p. 134). The meaning of compound sentences cannot be computed by looking at the semantic value of their components only, i.e., IS is *non-compositional*. Nonetheless, the meaning of compound sentences is still computable, their semantic value being *recursively* determined by the values of—many—less complex sentences.

7.1.2 The role of counterfactuals

Before interpreting the meaning of the causal modality, it is important to clarify the nature of the relation between causality and counterfactuals. This will allow me to reinterpret in an inferentialist perspective the failure of accounts that try to explicate the meaning of causality in terms of one or the other privileged inference—not only the counterfactual account (this section), but also other accounts (§7.2.1).

Recall that incompatibility entailments, which occupy a fundamental role in IS, are a sort of counterfactual claims. Yet, this does not mean that they can be used to reduce causality to counterfactual dependence. Very shortly put, this is because counterfactuals of the form 'If one were committed to p, one would not be entitled to q' belong to the metavocabulary, not to some object-language vocabulary with which 'causes' stands in a reducibility relation. Let me explain.

In IS, two sentences have identical meaning iff one can be substituted for the other *salva inferentia* (Brandom, 2000, chap. 4)—in other words, iff they have the same inferential role across all possible contexts, which *is* a matter of counterfactual dependences. Take two sentences p and q. They have the same meaning iff: if p (resp. q) follows from some inferential base,

[102] For an analogous remark with regard to the relation between inferentialist semantics and classical model-theoretic or possible-worlds semantics, see Peregrin (2008, pp. 100-101).

so does q (resp. p), and if p (resp. q) grants, together with certain collateral commitments, some inferential target, so does q (resp. p), together with the same collateral commitments. In terms of incompatibilities: p and q have the same meaning iff they are incompatibility-*equivalent*, viz. were one to endorse something that incompatibility-entails p (resp. q), one could not deny q (resp. p); and were one to endorse p (resp. q) and have certain collateral commitments, one could not deny whatever is incompatibility-entailed by q (resp. p) in conjunction with the same collateral commitments.

Now, both IS and Lewis' possible-worlds semantics (PWS) analyse the meaning of modal claims, causal claims included, in terms of counterfactually robust inferences. Yet, they do so in different ways.

In PWS, establishing the meaning of 'A necessitates (or raises the probability of, or causes, or...) B' goes via establishing truth conditions: the actual world is an A- and B-world, and the closest non-A world is a non-B world. This presupposes the possibility of evaluating counterfactual robustness in a way which is not theory- or language-relative, by reference to the intrinsic nature of A and B (no matter how they are described) and what happens at this and other worlds. In particular, for all possible A's and B's, 'causality' is—for Lewis—*reducible* to '(transitive) counterfactual dependence'.

In IS, instead, counterfactual robustness has to do with correctness of material inferences. This is a *normative* matter, which depends on context-dependent, collateral commitments, and the way they affect one's entitlements: whether B ought to be inferred from A depends on whether the rules of the game are such that other speakers have reasons to socially sanction one if one claims that A and also claims that, or acts as if, non-B.[103] The evaluation of such inferences cannot but be theory- and language-relative. In fact, the number and kind of collateral commitments may vary depending not only on *epistemological* factors, such as the availability of some type of evidence in a given context, but also on *semantic* factors, such as expansions of the theory or language frame. Due to non-monotonicity, any change of the latter kind results in a change in the counterfactuals on which meaning, holistically, depends. Whilst epistemological reasons make it difficult in practice to evaluate all counterfactuals involved, semantic reasons make

[103]Notice that correctness is not *reducible* to the fact that one is, on a given occasion, sanctioned or not. Rather, it is a consequence of the scorekeeping practice as a whole: a move is correct if it contributes to the 'harmony' of the practice, and wrong otherwise (see §8.1.1).

it impossible in principle to evaluate all possible counterfactuals.[104] Thus, meaning, which is *made explicit* in terms of commitments and entitlements, and the counterfactuals they induce, is not thereby *reducible* to them. Still, a change in theory/language need not *ipso facto* entail a substantial change in meaning. In fact, not all counterfactuals may be equally relevant. Meaning depends on *ranges* of counterfactual robustness (Brandom, 2008a, p. 108): two sentences share the same meaning *to the extent that* they have the same inferential role in a range of contexts, that is, to the extent that they meet the aforementioned, counterfactual criterion for the identity of meaning.

How is the criterion applied to the meaning of *causal* claims? Counterfactuals help assess to what extent the inferential roles of 'A causes B' and some other linguistic expression X (e.g., 'transitive counterfactual dependence') overlap. To be sure, counterfactuals may also be used to assess whether 'A causes B' and X have the *same* meaning (e.g., whether 'causality' is 'transitive counterfactual dependence'). But doing so results in undermining rather than supporting reductive analyses such as the counterfactual account, as I now turn to explain.

Let us extend the above identity criterion from sentences to *sets* of sentences. Two sets C and X have the same meaning iff they have the same semantic interpretant:

$$C \equiv_{\text{Inc}} X \text{ iff } \forall Y[Y \subseteq I(C) \leftrightarrow Y \subseteq I(X)] \tag{7.3}$$

where "$\equiv_{\text{Inc}}$" indicates incompatibility-*equivalence*. Next, let "C" stand for 'A causes B', and "A" and "$I(A)$" (resp. "B" and "$I(B)$") for 'A obtains' and 'A does not obtain' (resp. 'B obtains' and 'B does not obtain'). Since C is about some modal relation between the A's and the B's, the following condition must hold for X to count as conceptually equivalent to C:

$$C \equiv_{\text{Inc}} X \text{ iff } \forall A \forall B[A \cap I(B) \subseteq I(C) \leftrightarrow A \cap I(B) \subseteq I(X)] \tag{7.4}$$

In other words, 'A causes B' is incompatibility-equivalent to X iff 'A and non-B' is incompatible with 'A causes B' in all and only the cases in which 'A and non-B' is incompatible with X. In terms of commitments and entitlements:

[104] Accordingly, the use of the *ceteris paribus* clause among one's collateral commitments is to be understood as the acknowledgement that counterfactual statements have potential defeaters, not as an attempt to survey all possible defeaters (Brandom, 2008a, p. 107).

C is incompatibility-equivalent to X iff, were one to be committed to C one would not be entitled to 'A and non-B' iff were one to be committed to X one would not be entitled to 'A and non-B'. This amounts to a sort of normative adequacy condition for X to count as a reductive analysis of C.

Now, let us substitute 'B counterfactually depends on A' for 'X'. What we get is not the counterfactual account itself, but a statement on what it would take for the counterfactual account to be generally correct. However, some reflection shows that the counterfactual account is *not* generally correct, so doesn't exhaust the meaning of causality. As evidenced by the counterexamples in §4.2, there are relata and (this-wordly) scenarios such that commitment to the causal claim does not entail entitlement to the counterfactual dependence claim (e.g., overdetermination), and vice versa, commitment to the counterfactual dependence claim does not entail entitlement to the causal claim (e.g., chancy preemption). So, 'is caused by' and 'counterfactually depends on' do not have identical meaning.

Importantly, the failure of the counterfactual account is shown *from within* IS itself, in terms of incompatibilities not in turn interpreted in terms of truth. In the above analysis, meaning is not analysed descriptively, in terms of direct relations between vocabularies, but normatively, in terms of pragmatically mediated relations, which go from vocabulary to vocabulary through a set of practices that the mastery of language institutes. The incompatibility is not between claims as such (true or false) but between normative attitudes towards the claims (endorsed or not endorsed). Whether the counterfactual account ought to be accepted doesn't depend on whether it represents the world as it is, but on the speakers' ability to avoid social sanctioning by providing reasons when challenged. The speakers' attitudes towards the relation between 'facts' (e.g., their causal intuitions) and the correctness of linguistic practices (e.g., the way 'causality' as well as other concepts ought to be used) is key to demonstrate the adequacy of an explication. So, from the inferentialist point of view, the issue is not that the counterfactual account is *false*, period; rather, it is an example of an analysis that is *inappropriate* because non-pragmatically-mediated.

By applying the same strategy we can pinpoint the problems of other accounts that rely, in a way or another, on the isolation of some privileged inference to elucidate the meaning of causality (see §7.2.1 and §7.2.2).

7.2 The meaning of 'causes'

An inferentialist account of causality should help address the following two issues: (i) what kind of inferences causal claims license; and (ii) in virtue of what. The first issue can be—partly—addressed by discussing what class of meaning-constitutive inferences can make explicit, in the face of holism, the meaning of 'causes' (this section).[105] However, a satisfying account must also provide a *justification* for employing a privileged class of inferences to decide whether or not a causal claim ought to be endorsed. If holism can at most be contained not eliminated, an analysis that makes meaning explicit may be informative but never exhaustive. How can one make sense, in spite of this, of the objectivity of causal claims? I make some general points in this section and leave the discussion of the objectivity of causal claims to chapter 8.

7.2.1 Setting the stage

All claims license inferences. How can one distinguish *causal* claims from *non*-causal ones? What inferences are constitutive of the meaning of causal claims? The most prominent accounts of causality have been reviewed and criticised already in chapters 4–6. Here I discuss why the accounts that rely on the identification of some privileged inference are inadequate *by the inferentialist's own light*. This will help me point to the kind of analysis that causal claims require—which is the analysis I offer in §7.2.3.

Difference-making accounts tend to reduce the meaning of causal claims to truth conditions, which in turn are identified with some privileged test condition, such that the satisfaction of the condition is meant to exhaustively characterise causality. However, no such condition seems to exist. As argued, causality has excess content with respect to difference-making criteria (see §6.3). Now we are in the position to rephrase the excess content thesis in IS terms. Let us indicate with "C" and "X", for *whatever* couple of relata A and

[105] It must be pointed out that the Brandomian, *normative*-functionalist variety of inferentialism that I favour differs from *causal*-functionalist varieties, such as the folk theory of causation mentioned in fn.48. According to the former, strictly speaking only inferential *rules* not inferences can be meaning-constitutive. In contrast, the latter focus on the causal role of expressions, viz. the causal relations that underwrite—actual—inferences to and from the expressions (cf. Peregrin, 2012, fn. 2). Causal-functionalist analyses of causality may be subject to the following objection: if functional role is *causal* role, as functionalist theories of mind have it, a theory of 'causes' is necessarily circular. Normative-functionalist varieties, instead, are immune to such an objection: meaning has to do with the inferences that *ought to be* drawn, not those that are *actually* drawn. Reference to actual inferences is only envisaged as useful to model, or make explicit, the underlying inference rules (cf. fn. 102).

B, respectively 'A causes B' and 'The relation between A and B satisfies test condition X'. Difference-making accounts (DM) are committed to something along the following lines:

[DM] $C \equiv_{\text{Inc}} X$

Such an incompatibility equivalence translates 'C iff X' into the IS jargon: if one were to be committed to X, one would be committed to C, and—vice versa—if one were to be committed to C, one would be committed to X. But the counterexamples in chapter 4 show that there is no X which satisfies the equivalence for *all* causal claims. So, if anything, this should teach us that formulating the analysis in incompatibility-equivalence terms is not a promising route. We need something weaker than that.

Other analyses either emphasise the role of base-to-claim inferences (BC) and downplay the role of claim-to-target inferences (CT), or vice versa. I believe that Reiss (2011)' pluralist account represents well the former attitude, whereas Price (1998)'s expressivist position represents well the latter. Since, like me, neither of them is in the business of providing truth conditions, but rather of accounting for meaning as *use*, I will consider their proposals in terms for how well they accommodate the meaning of causal claims in terms of their *inferential role*. Before doing this, however, I need to introduce the framework in which my analysis will be conducted.

Recall that IS takes as the content of a sentence the set of sentences that express propositions materially *incompatible* with it (cf. §7.1.1). Thus, strictly speaking, the various proposals should be rephrased in terms of sentences that belong to the *incompatibility set* of the target sentence, viz. the sentences with which the target sentence stands in a relation of primitive incompatibility. How can one get a grip on what the incompatibility set of a causal claim is? Here is an idea: make explicit the content of the claim in terms of the commitments that 'causes' institute, their relation with entitlements and lack thereof. This can be done by analysing the meaning of the causal claim in terms of incompatibility entailments—rather than primitive incompatibilities—between the claim and other claims. And since a causal claim is generally compatible or incompatible with other claims only in conjunction with other, collateral commitments, the relevant incompatibility entailments will be *multi-premiss* arguments, where the conclusion follows

only in the presence of a complex set of commitments. More precisely, a typical analysis will be in terms of the tuples of sentences whose elements belong to the causal claim's *inferential potential* (cf. Peregrin, 2008, §3-§4).

From here onwards, let us denote with "c" a *specific* causal claim (e.g., '*Smoking* causes *lung cancer*'), relative to a specific couple of relata $< A, B >$ (e.g., <smoking, lung cancer>), as opposed to the class C of *all* causal claims; and let us denote with "$C_{<A,B>}$" the class of the c's relative to $< A, B >$, i.e., 'causes' relative to $< A, B >$ (e.g., '$C_{<\text{smoking, lung cancer}>}$'). The inferential potential of a sentence c, in short c^{ip}, can be defined as the set of c's downstream potential, in short $c^{\downarrow}$, and upstream potential, in short $c^{\uparrow}$:

$$c^{ip} = \{c^{\downarrow}, c^{\uparrow}\}^{106} \tag{7.5}$$

The downstream potential of c is the set of sentences p from which c follows:

$$c^{\downarrow} = \{p \mid p \vDash_{\text{Inc}} c\} \tag{7.6}$$

The upstream potential of c is the set of tuples comprising the sentences r which follow from c and the collateral premisses q which aid the inference from c to r:

$$c^{\uparrow} = \{< q, r > \mid q, c \vDash_{\text{Inc}} r\} \tag{7.7}$$

A typical analysis, or conceptual explication, will have $C_{<A,B>}$ as the *explicandum*, and the tuples comprising the downstream and upstream potential of c, that is the *contexts* (or context descriptions) that fix the correct use of c, as the *explicans*:[107]

$$C_{<A,B>} = \{< p, q, r > \mid p \in c^{\downarrow}, < q, r > \in c^{\uparrow}\} \tag{7.8}$$

Now, let us bear in mind that the context descriptions such that c is *inferrable* from p and *grants* inference to some r_k $(k = 1, \ldots, w)$ are typically *conjunctions* of sentences not single sentences. Accordingly, the analysis should be modified to allow for multi-premiss arguments where c and r_k follow from conjunctions of, respectively, p's in $c^{\downarrow}$ and q's in $c^{\uparrow}$—in short $\wedge p_i$ $(i = 1, \ldots, u)$ and $\wedge q_j$ $(j = 1, \ldots, v)$:

[106] From here onwards, I will introduce downstream and upstream potential by using the labels "$\boxed{c^{\downarrow}}$" and "$\boxed{c^{\uparrow}}$".

[107] "C" stands here for a particular (binary) predicate. However, the inferentialist analysis schema is in principle the same for all concepts.

$$C_{<A,B>} = \{< \wedge p_i, \wedge q_j, r_k > \mid \wedge p_i \in c^{\downarrow}, < \wedge q_j, r_k > \in c^{\uparrow}\} \qquad (7.9)$$

Notice that the *explicans* comprises occurrences of "c", which in turn comprise occurrences of "$C_{<A,B>}$". However, since c is different from $C_{<A,B>}$, no circularity arises. In line with the inferentialist strategy to explicate the meaning of subsentential locution in terms of the meaning of sentences (not vice versa), here the meaning of '$C_{<A,B>}$', viz. the predicate 'causes' as occurring in c, is explicated in terms of *more basic* incompatibilities between c and other sentences. Obviously, evaluating such relations presupposes a grasp, or mastery, of $C_{<A,B>}$ on the part of the participants in the language game. But this is besides the point: one thing is to grasp, or master, the meaning of a linguistic expression; another thing is to make it explicit.

Also notice that the meaning of '$C_{<A,B>}$' is fixed by all possible incompatibilities which underlie the correct use of c, including those that are not transparent to the speakers, but to which they are nonetheless bound by the rules that their normative attitudes institute. No explication will be exhaustive unless the totality of inferentially connected sentences, both actual and potential, are identified. But this is not a problem for my project, which has no reductive pretensions. My task, in fact, is to *make explicit* meaning in terms of commitments and entitlements which are acknowledged as relevant, not to *reduce* meaning to such commitments and entitlements.

7.2.2 Base vs target?

Back to Reiss and Price. Reiss first. For Reiss, the meaning of a causal claim is given by its inferential connections with other propositions, and its method of verification determines what these inferential connections are (see Reiss, 2011, p. 923). Verification conditions, although not providing an *exhaustive* analysis of 'causes', still largely determine what 'causes' means (Reiss, 2012, §3.1). The most straightfoward interpretation of Reiss' view is that the meaning of each causal claim is fixed by its verification method, which naturally leads to the view that there are at least as many distinct meanings as there are methods of verification. Each method can be identified with the inferences actually drawn by some community (K), on the basis of their theoretical background and the evidence available to them (X). What counts as evidence here clearly depends on K's standards of verification. The consequences of c, viz. $c^{\uparrow}$, will be limited to the sort of circumstances of c's appropriate application, viz. $c^{\downarrow}$,

in line with the fragmentation of purposes suggested in (Reiss, 2009a, 2011). Very roughly put, this means that if c was 'retrospectively' established in the context of providing an explanation, the consequences of its appropriate application will be other explanations (E), in analogous contexts; if c was established in the context of predicting possible effects of putative causes, the consequences of its appropriate application will be predictions (P), in analogous contexts; and if c was established in the context of an intervention, the consequences of its appropriate application will be claims about the result of possible interventions (I), in analogous contexts.

How should this translate into IS terms? Let us indicate conjunctions of base sentences with "$\wedge x_i$" ($i = 1, \ldots, m$) and target sentences with "e_j" ($j = 1, \ldots, n$). Here is one possible way to cash out BC relative to K, corresponding to the case where the target is E:

$$[\mathbf{BC}_{K,E}] \quad C_{<A,B>} = \{< \wedge x_i, e_j > \mid \wedge x_i \in c^{\downarrow}, < \wedge x_i, e_j > \in c^{\uparrow}\}$$

That is, the meaning of $C_{<A,B>}$ is made explicit by the pairs $< \wedge x_i, e_j >$, such that if a member of K were to be committed to $\wedge x_i$, one would be committed to c as well as the sort of claims that (according to K) $\wedge x_i$ allows in the BC context, in this case e_j. Notice that, if the meaning of $C_{<A,B>}$ is—literally—the method of its verification, arguably $\wedge x_i$ *and* c is equivalent to $\wedge x_i$. As a result, one may correctly infer e_j from $\wedge x_i$ only, and drop c from the premisses without any loss. Also notice that the method of verification could as well have entitled one to P or I, in which cases the corresponding analysis would have been, respectively, $\mathrm{BC}_{K,P}$ and $\mathrm{BC}_{K,I}$. But in line with Reiss' idea that purposes disunify meaning, for any method arguably only one amongst E, P, and I is the appropriate target. Now, $\mathrm{BC}_{K,E}$ (respectively, $\mathrm{BC}_{K,P}$ or $\mathrm{BC}_{K,I}$) provides the correct analysis if the meaning of $C_{<A,B>}$ is—nothing but—the method of its verification. Two problems arise here.

On the one hand, the first clause of $\mathrm{BC}_{K,E}$ is too strong: it does not allow for the possibility that K is mistaken, and beliefs need revising in the light of new commitments. For instance, K could think that the relation is causal on the basis of one method, but the method is not suited to spot spurious relations in the context in which evidence was gathered. In general, K remain committed to a set kind of consequences, by the light of their own verification standards, no matter what external input comes in, from other communities,

their different background knowledge, and their different methods to establish the claim. The problem is that the identity between meaning and verification places too strict a constraint on meaning. To solve the problem, one would need to make room for the possibility that a larger variety of evidence is allowed. But if meaning is verification, enlarging the inferential base beyond X, that is, the evidence acknowledged as relevant by K, would render the material inference automatically incorrect.

On the other hand, the second clause of $\mathrm{BC}_{K,E}$ is too weak: the analysis limits the applicability of c to the sort of claims allowed by X, namely the claims that are appropriate to the context in which c was first established. For instance, if $\wedge x_i$ describes the context of an RCT on mice, from $\wedge x_i$ it follows only that some net positive effect will be observed in a population which is as similar as possible to that of the experiment (e.g., a randomised population of mice, not a cohort of humans). So, K may think that c is only safely applicable to base contexts allowed by $\wedge x_i$, whereas in fact it has the *potential* to carry over to different target contexts, populations, etc., based on studies carried out by other communities, by means of other methods, and more generally on other theories and hypotheses. This is, once again, the excess content thesis: the meaning of (many) causal claims is not reducible to their inferential base, or test conditions, i.e. features of the context of their appropriate application.

Now, to the extent that Reiss' aim is not to reduce $C_{<A,B>}$ to $\wedge x_i$, he seems to implicitly acknowledge this. So, one may interpret him as saying that, although $\wedge x_i$ largely determine $C_{<A,B>}$, $C_{<A,B>}$ has excess content with respect to $\wedge x_i$, and $\wedge x_i \& c$ *together* may have broader consequences than $\wedge x_i$ *alone*. In IS terms: something incompatible with the target may be incompatible with $\wedge x_i$ without being incompatible with c itself. Still, the fact remains that the only *kind* of consequences of c are those that follow in the BC context, whereas the analysis should have allowed for the target's scope being broader. But to the extent that meaning is tied to verification conditions, the inference target, too, cannot be enlarged: no consequences can be drawn from c that weren't actually verified according to K's standards. Instead, if one allows collateral commitments to be meaning-constitutive so as to enlarge the target of $C_{<A,B>}$, one thereby abandons the view that meaning and verification method are (so) strictly related—and one comes closer to the sort of inferentialism I have in mind.

An alternative is to go the opposite direction, viz. emphasise CT (use

conditions) and downplay BC (test conditions). This is what Price does. For Price, causality is one of those concepts for which only 'usage conditions' can be provided (cf. §4.1, fn. 28), along the following, *expressivist* lines: the utterance "X is R" is *prima facie* appropriate when used by a speaker who experiences response R in the presence of X, that is, when he is in some psychological state ϕ (see Price, 1998, §2). More precisely, for Price c is granted by a particular response experienced by the agent, namely his being in the psychological state of believing 'I can freely manipulate A to effectively bring about B' (R). And arguably, for Price, the consequences of the appropriate use of c are claims about the results of interventions, i.e., claims of the form 'Intervening on A is an effective means to bring about B' (I). How can we put this into IS terms?

We may interpret 'being in the psychological state ϕ' as sufficient to 'being committed to c'. In other words, the endorsement of c is the subject's own way to respond to ϕ by committing himself to the belief that c.[108] That this reading is in line with Price's expressivism is suggested by his remark that the usage condition provides *subjective* assertibility conditions (see Price, 1998, §2). Analogously, we could interpret the psychological state of believing in the causal relation as binding one to I. As a result, the response-dependent analysis of causality can be cashed out as follows:

$$\textbf{[CT]} \quad C_{<A,B>} = \{< \wedge r_i, i_j > \mid \wedge r_i \in c^{\downarrow}, < \wedge r_i, i_j > \in c^{\uparrow}\}$$

where "$\wedge r_i$" ($i = 1, \ldots, m$) stands for conjunctions of base sentences and "i_j" ($j = 1, \ldots, n$) stands for target sentences.

But the analysis is wanting. To begin with, the BC step relies on a sort of *psychological* response, or experience. But this does not constitute a satisfying explication of the *rational* process of acceptance of scientific hypotheses.[109] Moreover, the analysis does not explicate properly the CT step, either. As said in §4.1, the agency account suffers from the problem that many causal

[108] Commitment to the response does not presuppose that the subject be aware of his response, only that he ought to commit to the claim 'I have (had) the response', *if* he were made aware of it.

[109] Menzies and Price (1993) *do* offer an analysis of causality in terms of manipulability and resulting probability raising relations; however, according to the expressivist reading of the agency theory, the notions of manipulability and probability raising enter the picture only to provide use conditions, not test conditions (by which the claim is rationally endorsed).

relations are not manipulable. Here, c commits one to i_j, so the analysis is inadequate. Take the causal claim 'Friction between continental plates causes the earthquake'. This is supposed to incompatibility-entail 'Intervening on the friction between continental plates is an effective means to affect the earthquake'. However, there being no way to effectively manipulate the plates, which is incompatible with *me* being able to affect the earthquake by manipulating the plates, is not incompatible with *friction* causing the earthquake. There is more to (the upstream potential of) causal claims than the possibility of effective manipulation, namely all those objective facts that follow from the claim but do not involve facts about agency. More generally, the inferential target is too narrowly conceived. Besides I, other consequences, namely E and P, follow from c. So the analysis is at best incomplete.

The usage condition seems to miss a crucial aspect of the meaning of causal claims, viz. their objective dimension. Aren't there objective assertibility conditions, too, besides subjective assertibility conditions? Not only is Price's expressivism deflationary about reference and truth conditions, it offers no tool to interpret our discussion about them as rational. A good analysis should account for the fact that we regard causal language as referential and causal claims as true or false.

A final consideration should be made as regards both of the above analyses. A crucial point that hasn't emerged so far, but is implicitly presupposed by the inferentialist approach, is that the target of a sentence comprises *all* the claims that the sentence in question contributes to warrant. In the case of causal claims, this means that the CT step should not be limited to inferences where claims involving *one* relatum are granted (i.e., explained or predicted, in actual or counterfactual circumstances) by claims involving the *other* relatum, together with the causal claim and collateral premisses. It should comprise *all* the inferences that the causal claim contributes to warrant. The causal claim may—and typically does—entitle one to *other* inferences, namely inferences to claims not involving one or the other relatum. To make a trivial example, 'smoking causes lung cancer' entitles one not only to infer lung cancer in the presence of smoking, certain hereditary features, etc., but also to infer shorter life expectation, when the collateral commitment 'lung cancer shortens life expectation' is added to the premisses.

7.2.3 The meaning of causal claims

Let me recap the argument so far. Following Sellars' observation that 'causes' is an inference license governing the use of the relata, I've suggested that the meaning of 'causes' should not be analysed in terms of the contribution of 'causes' to the truth conditions of causal claims, but in terms of the inferential role of the sentences in which 'causes' appears. To put it in a slogan: the meaning of 'causes' depends on its *contribution to the correctness of arguments* involving the use of one relatum or both relata. (How this contribution is to be characterised is an issue I leave to chapter 8.)

Let me also briefly recap the main points that have emerged from §7.2.2: (1) incompatibility-equivalence analyses are inadequate; (2) the inferential base should not be limited to information made *available* on the basis of one verification method; (3) it should be in principle possible for the target to be broader than the set of sentences which have been used to support c, and which c, if successfully employed, would in turn support; (4) the target should contain not just claims about interventions but also claims about predictions and explanations; (5) the target should comprise *all* the claims that the causal claim contributes to warrant. Accommodating 1 is straightforward.

A way to deal with 2 and 3 at once is to prevent that the target be limited to the sort of consequences that follow from the base X, by allowing collateral commitments Y, based on other evidence and theoretical knowledge, to play a role in CT. As a result, the target can generate extra evidence and theoretical knowledge, which may then become part of X. The reason for distinguishing X and Y is that the commitments that are relevant to establishing c and those that are relevant to using it need not be the same. Often, in fact, the contexts of correct application of c itself (laboratory experiments, RCTs, and controlled environments in general) are different from the contexts of correct application of its consequences (natural, or non-controlled, environments). As Cartwright would put it, a good theory for "hunting" causal claims is not *ipso facto* a good theory for "using" them, and vice versa (see Cartwright, 2007b, pp. 48-49). So, a good analysis ought not collapse the two. In particular, it seems reasonable to impose the following two constraints, namely that Y be—minimally—(i) *compatible* with X but *less restrictive* than X (for the causal claim to have wider applicability than its base), and (ii) compatible with c itself. These constraints are intuitively met if $X \vDash_{\text{Inc}} Y$ (figure 7.1).

To deal with 4 and 5, we should consider *all* the claims that c licenses. In

fact, the meaning of c is determined by *all* its possible uses. Accordingly, c must be analysed in terms of *all* the inferences that ensue from it.[110] Among them are, minimally, inferences that on the basis of knowledge of actual or possible causes aim to—respectively—predict actual effects (actual scenarios) and possible effects (counterfactual scenarios); and there are inferences that on the basis of knowledge of actual or possible effects aim to retrodict, respectively, actual causes (actual scenarios) and possible causes (counterfactual scenarios), thereby providing causal *explanations*.[111] This ensures that both directions of inference, viz. cause-to-effect and effect-to-cause, are considered. Furthermore, as pointed out at the end of §7.2.2, the claim helps warrant other claims as well. I will label with "Z" the set of *all* consequences that follow from the causal claim.

Let us indicate conjunctions of base sentences with "$\wedge x_i$" ($i = 1, \ldots, u$), conjunctions of collateral commitments with "$\wedge y_j$" ($j = 1, \ldots, v$), and target sentences with "z_k" ($k = 1, \ldots, w$). What follows is a proposal which spells out more formally the view first introduced in §6.5. The meaning of claims of the form 'A causes B' (relativised to *specific* relata A and B) is—minimally—characterised in terms of their inferential potential (INF):

$$\textbf{[INF]} \quad C_{<A,B>} = \{< \wedge x_i, \wedge y_j, z_k > \mid \wedge\, x_i \in c^{\downarrow}, < \wedge q_j, z_k > \in c^{\uparrow}\}$$

According to INF, the meaning of '$C_{<A,B>}$' is made explicit in terms of the contribution of '$C_{<A,B>}$' to the correctness of the arguments in which "$C_{<A,B>}$" appears. More precisely, '$C_{<A,B>}$' is made explicit by the tuples $< \wedge x_i, \wedge y_j, z_k >$, such that $\wedge x_i$ commits one to the application of c, and $\wedge y_j$ entitles one to use c to infer z_k. INF satisfies the intuition that causal claims may have broader or narrower scope (figure 7.1). This depends on the (meaning of the) base, or test conditions X, and the relation between collateral commitments, or use conditions Y, and consequences Z. The dynamics between commitments and entitlements are such as to allow for 'conservative' revisions of the meaning of $C_{<A,B>}$, by modification of $< \wedge x_i, \wedge y_j, z_k >$

[110] The exact form of target claims will depend, besides the commitment to c itself, on the Y that best suits a given context. More on this below and in chapter 8.

[111] Notice that inferring the cause given the effect amounts to an ætiological explanation of the effect in terms of the cause. From here onwards, I take it that *retrodiction*, or ætiological explanation, is the sort of explanation that causal relations entitle one to. This excludes other sorts of explanations, e.g., teleological, constitutive, explanation by unification, etc.

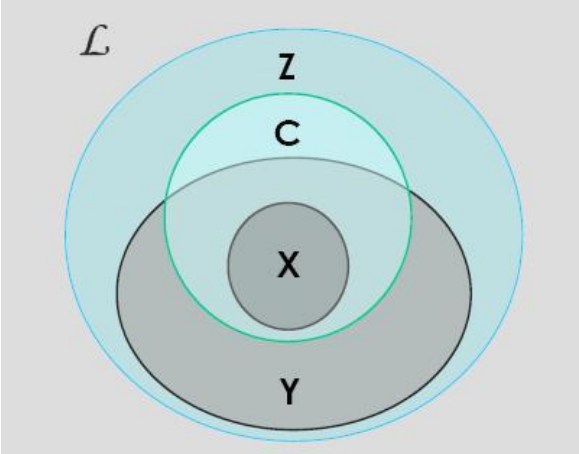

Figure 7.1: (In)compatibility relations among C, X, Y and Z. $I(C)$ belongs to $I(X)$ and overlaps with $I(Y)$. $I(Y)$ belongs to $I(X)$. $I(Z)$ belongs to $I(C \cup Y)$. The relations among C, X, Y and Z may change, depending on changes in commitments and entitlements, which in turn depend on changes in beliefs and/or changes in the language $\mathcal{L}$.

(more on this below). The vagueness of 'causes' is illustrated in terms of the sets' boundaries being fuzzy (cf. §6.4), which in turn depends on borderline material inferences being neither definitely correct not definitely incorrect.

I will now illustrate INF by reference to a claim involving a cause of apoptosis, 'XIAP feedback promotes irreversibility of Casp3 activation', in short c_1, as discussed in (Legewie et al., 2006). Here INF should deliver an explication of '$C_{<\text{XIAP feedback, irreversibility}>}$'. Background knowledge is used to identify relevant interactions in the intrinsic pathway after release of cytochrome c from the mitochondrion: Apaf1's activation of Casp9, Casp9's cleavage of Casp3, Casp3's cleavage of Casp9, XIAP's inhibition of Casp3 and Casp9 (figure 3.1). Quantitative study of the kinetics by means of ODEs strongly commit to the claim: in cases of bistability, the feedback induced by XIAP's competitive binding is necessary for irreversibility (BC step; see figure 3.3). More precisely, the following inference is part of '$C_{<\text{XIAP feedback, irreversibility}>}$':

$\boxed{\text{c}^{\downarrow}}$ given Casp3 feedback, XIAP feedback is regularly associated with irreversibility; irreversibility counterfactually depends on XIAP feedback; switching off XIAP feedback affects irreversibility; ... $\vDash_{\text{Inc}} c_1$

In turn, adding the claim to suitably chosen sets of premisses entitles to narrow-scope, precise inferences as regards the possible way in which Casp3

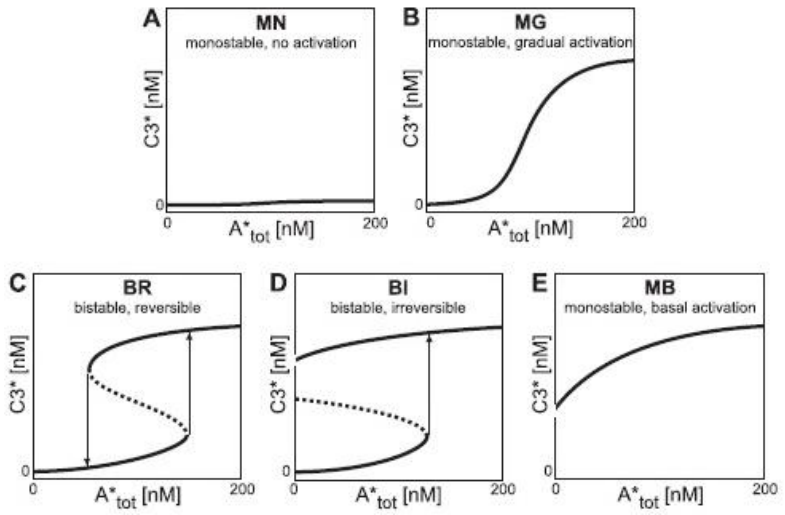

Figure 7.2: Dependence of different qualitative behaviours on the presence of different species and concentrations (from Legewie et al., 2006, p. 1067).

activation may obtain in the cell (figure 7.2). Among such inferences are explanations as well as predictions, whether about the actual course of events or about the result of interventions. In both cases, the inferences do not necessarily comprise claims regarding the obtaining of one relatum as the conclusion. Below are examples of prediction.

$\boxed{c^{\uparrow}}$ c_1, XIAP competitive binding, Casp3 feedback, XIAP & Casp9 > Casp3 $\models_{\text{Inc}}$ bistability and irreversibility (**BI**); c_1, Casp9 > XIAP & Casp3 $\models_{\text{Inc}}$ monostability & gradual activation (**MG**)

Notice that information regarding the relative concentrations of the reactants, which is included in the collateral premisses and was not present in the inferential base, is crucial for the correctness of the predictions.

Examples of explanation are:

$\boxed{c^{\uparrow}}$ c_1, bistability & irreversibility (**BI**) $\models_{\text{Inc}}$ XIAP competitive binding; c_1, bistability & reversibility (**BR**) $\models_{\text{Inc}}$ no Casp3 feedback; c_1, monostability & no activation (**MN**) $\models_{\text{Inc}}$ XIAP > Caspases

A number of issues deserve some discussion. Does INF commit one to the view that the meaning of $C_{<A,B>}$ reduces to regularities between antecedents and consequents? That is, does INF reduce to a regularity view of causality?

No. INF makes explicit '$C_{<A,B>}$' in terms of commitments leading to entitlements or lack thereof. The relation between commitments and entitlements is often such as to produce regular enough associations between endorsement of antecedents and (successful) endorsement of consequents. Yet INF allows that in some contexts, e.g., indeterministic scenarios, chaotic or context-sensitive scenarios, etc. material inferences turn out incorrect more often than in other contexts. This is because the boundaries between the meaning-constitutive sets are fuzzy. In borderline cases, one should expect no regular association between antecedents and consequents. Yet, it remains rational to have a (less-than-maximal) commitment to the consequent given the antecedent. What inferences are—in spite of this—meaning-constitutive depends on a balance between simplicity and fruitfulness of our commitments.

Need a causal claim always entitle one to the *same* kind of claims that warranted its correct application? No. For instance, one may establish a causal claim on the basis of an intervention. However, the intervention may modify the context so that one cannot infer from the claim that a similar intervention will produce a similar effect. So, there is a sort of asymmetry between test and use conditions. This shows that 'causes' is not always 'harmonious' in Dummett's sense. For Dummett (1991, chap. 9), harmonious concepts are such that introduction (BC) and elimination (CT) rules are function of one another. Rule modifications shouldn't entitle one to draw conclusions that were not warranted by the methods of arriving at the premisses. In formal theories, the addition of predicates, functors, axioms, etc. may (and typically does) make it possible to prove statements one could not express in the old theory. But this should *not* make it possible to prove statements expressible in the original vocabulary that were not provable in the original theory. This reasoning, for Dummett, should extend to natural languages: additions, or modifications (e.g., weakening of introduction rules, or strengthening of elimination rules), of expressions E are inadmissible if E make it possible to derive consequences expressible in the old language that could not have been derived otherwise, without E. However, it is not clear that in natural languages harmony should be interpreted in this way. For instance, in science it is common that the addition of theoretical terms allows one to draw new conclusions expressible in the old language, but this merely shows that the new terms carry substantive content, not that there is something wrong with their rules of application (see Brandom, 2000, p. 71). There are reasons to believe that 'causes' works more like a theoretical term than a logical term.

Whereas logical concepts, e.g. 'or', are definable in terms of introduction and elimination rules such that side assumptions (e.g., collateral commitments in Y) that figure in them aren't meaning-constitutive, causality is a concept whose rules are less clear-cut, and whose content is fixed by a larger base of concepts and cannot be totally isolated from them. This is in line with the so-called 'theory-theory' of concepts (Laurence and Margolis, 1999, §4).[112]

Do explanations, predictions and interventions *necessarily* belong to the target? Whereas c_1 does entitle one to interventions (in a non-ideal sense), other causal claims don't, e.g., 'the gravitational attraction exerted by the moon causes tides of such-and-such a magnitude'. In that case, the counterfactual cause-to-effect inference cannot be interpreted as a claim about the result of a *physically possible* intervention. Nonetheless, it may still be the case that Y and c entitle us to the counterfactual claim, provided this depicts a scenario where the relation between antecedent and consequent fits particularly well with Y. For instance, the moon-tides system can be regarded as a limiting case, which stands in an analogical relation with other systems—whose behaviour depends on the same forces, same configuration of the component parts, etc.—where interventions *are* possible. So the moon-tides counterfactual may, if one wishes, be interpreted as a claim about the result of an intervention which is 'possible *by extension*'. However, it need not be so interpreted. The analogy among the systems in the class, which is itself identified in terms of similarity of inferential role, is what makes the counterfactual inference (materially) correct. Whether or not some scenario counts as a limiting case in the above sense depends on the content of our previous commitments, theories, etc. This is not an all-or-nothing affair, but rather a matter of degree: the weaker the analogies and intuitions, the weaker the entailments. One may interpret the analogy as justifying the assumption that the moon has such-and-such a capacity, or that it obeys such-and-such a law, or... and in turn the commitment to the stability of the capacity, or the law, or... as justifying the inference. In any case, what matters is not whether explanations, predictions and interventions belong to the target. "Explanation", "prediction" and "intervention" are themselves labels we attach to forward- and backward-looking, actual and counterfactual inferences. Their being explanations, predictions or interventions does not depend on anything intrinsic, but rather on the value they assume *for us*. Saying that a causal

[112] This view need not be in opposition to the view that 'causes' has a 'prototypical' content. More on this in §7.2.4.

claim qualifies as such only if its target comprises explanations, predictions and interventions is just shorthand for saying that a causal claim is an inference license that entitles one to correct forward- and backward-looking, actual and counterfactual inferences. The range of inferences in the target may actually vary from claim to claim.

Notice the difference with AG and INT. In AG, unmanipulable relations count as causal in virtue of the resemblance between their intrinsic properties and those of some analogous situation where interventions are possible. In INF, instead, the being causal of *all* relations, whether manipulable or not, depends on their being inference licenses. What matters is not similarity of intrinsic properties, but similarity of inferential role. In INT, a relation is causal in virtue of the existence of ideal interventions which would—counterfactually—result in a change in the effect. Relations which don't allow for *physically possible* interventions are causal iff ideal interventions are possible in a *logical or conceptual* sense. INF, instead, demands neither that causal claims be analysed in terms of 'ideal' interventions nor that there be a fact of the matter as to the correctness of our intuitions on what interventions count as logically or conceptually possible. In either case, although similarity of inferential role is ultimately judged in terms of closeness to the inferential role of some 'prototypical' notion of cause (see §7.2.4), which also allows for inferences about the result of interventions, nothing prevents that a relation which cannot be manipulated is closer to the prototype that one that can, provided other conditions are met, e.g., the claim can be inferred from more test criteria, it allows for better prediction and explanation, etc. Upshot: the meaning of causal claims is not reduced to manipulability relations, but analysed in terms of a cluster of criteria (one of which is manipulability) that all contribute to licensing certain inferences.

How can INF guarantee the requirement of causal asymmetry? Since causal relations are asymmetric, the content of '$C_{<A,B>}$' also contains a distinction between which of the two relata is the cause and which is the effect. According to INF, there is nothing intrinsic in A and B that makes one be the cause of the other, but not vice versa. On the contrary, both $C_{<A,B>}$ and $C_{<B,A>}$ may be correctly applicable, only to different contexts. Their correct applicability presupposes distinguishing forward-looking from backward-looking counterfactuals, which in turn presupposes assuming the direction of time as *given*. Which of the facts described by the premisses and conclusion come first and which come later belongs to the background knowledge of those

who endorse the inferences and commit themselves to the claims. This move is often blamed on the ground that it prevents the possibility to explain (non-circularly) the directionality of time as supervening on causal facts. However, one need not exclude the possibility that time *globally* supervenes on facts about causal relations, whilst at the same time maintaining that temporal information contributes *locally* to the meaning of a causal claim, relative to some specific context. In fact, in addition to background knowledge (e.g., other causal relations on which the target relation depends) and knowledge of the context (presence or absence of other causally relevant factors), one may need temporal cues (time of events, values of variables at specific times, etc.) to say what causes what, information that may be unnecessary in the case where knowledge of the entire state of the universe is available.

Finally, as I mentioned, INF allows for 'conservative' revisions of meaning. So one may wonder: what about *non-conservative* revisions? What happens if BC inferences or CT inferences are mistaken? One may think that all revisions fall neatly into either the conservative category or the non-conservative category. As a result, either we get a shift in meaning or we don't. This will seem uncontroversial to those who reduce meaning to truth conditions: either something bears on the truth conditions or it does not; whence, either it is relevant to meaning or it is not. If it does not bear on truth conditions, it leaves the relation between truth conditions and c unchanged; if it does, it either strengthens or weakens the relation, by contributing to spell out more clearly the truth conditions. However, those that are sympathetic to the idea that meaning is inferential role and is holistically fixed will be led to different considerations. Whenever one's language or theory changes, we get *some* shift in meaning (see Brandom (2008a, p. 108); cf. Harman (1999, pp. 134-137)). So, what really matters is not *whether* there is a shift, but *how relevant* the shift is. Conservative revisions result in a shift that does not shake the meaning of $C_{<A,B>}$: the relation remains causal, there is only a slight change in c's inferential potential. Non-conservative revisions, instead, result in a shift that leads to major changes in meaning. Two important issues now arise, and need to be distinguished. On the one hand, there is the issue of evaluating the *magnitude* of the shift in meaning, and in so doing, deciding whether or not the belief that a particular relation is causal should be amongst one's commitments. What makes for a minor shift and what makes for a major shift is discussed in chapter 8 in the context of providing a criterion for evaluating the objectivity of the causal relation. Just to anticipate, this will be

done by providing some general guidelines on how to assess the contribution of '$C_{<A,B>}$' to the correctness of the arguments in which "$C_{<A,B>}$" appears. On the other hand, there is the issue of identifying the conditions for a relation to count as 'causal'. If one rephrases the issue concerning the relations as an issue concerning the *predicates* describing such relations, the question becomes: Which predicates stand for *causal* inference licenses, and which stand for *non*-causal inference licenses? It is on this second issue that the rest of this section is focussed.

7.2.4 'Causes' *simpliciter*

INF constitutes an analysis of the meaning of causal claims of the general form 'A causes B', or, to put it better, an analysis schema, that can be filled differently depending on the claim in question. I will now turn to the issue of determining, given the meaning of the sentences in which 'causes' appears, the meaning, if any, of the subsentential locution 'causes' itself.

This issue is important for the following reason. So far, the leitmotif of this chapter has been that 'causes' should be interpreted as an inference license governing the use of the relata 'A' and 'B'. However, one may interpret *all* binary predicates as inference licenses of one sort or another. What makes a predicate belong to the class of *causal* predicates? What makes an inference license be a *causal* inference license? To answer one needs to characterise the meaning of 'causes' *simpliciter*, so as to distinguish how 'causes', as well as other 'causes'-like predicates, as opposed to *non*-'causes'-like predicates, contributes to the correctness of the arguments in which it appears. This is to allow one to decide whether or not a predicate belongs to the general category of 'causes'—with some caveats.

Let us assume that 'causes' itself is, by default, at the centre of the causal cluster. What else does belong to the cluster, and what doesn't? It is reasonable to expect from the criterion that relations such as 'contributes', 'inhibits', 'promotes', 'prevents', etc. turn out causal. At the same time, the criterion should exclude other relations from the cluster. To the non-causal category should surely belong semantic, logical or mathematical relations, such as those holding of the pairs <bachelor, unmarried man>, $< a$, non-$a >$, <cat, mammal>, <mean, variance>, etc. Among the non-causal predicates one also expects to find predicates describing empirical relations that have very little, or nothing, to do with causality, e.g., 'to the right of', 'higher

than', etc. What is *not* reasonable to expect is that the criterion always deliver yes-no verdicts. If causal content is measured in terms of inferential role, 'being causal' may not be a yes-no issue, but a matter of degree. A predicate may have causal content in one context, not in others. How much causal content it has depends on how similar it is to the prototype 'causes', viz. on the extent to which one can be substituted for the other *salva inferentia*. So, the question becomes: *How close* is the license to the centre of the cluster?

The final product of the analysis will in this case be a set of 'canonical' introduction (BC) and elimination (CT) rules, in line with the so-called 'prototype' theory of concepts (see Laurence and Margolis, 1999, §3). This is not to claim, along with (Dummett, 1991, chap. 10), that 'indirect' ways of introducing or eliminating causal claims are meaning-constitutive *only if reducible* to 'direct', canonical rules. Although, as a matter of fact, such a reduction is often possible, typically for instances of C that lie closer to the centre of the cluster, it need not always be so. For each token concept $C_{<A,B>}$, the boundaries of the sets in INF are fuzzy, corresponding to cases where correct application cannot be decided based on clear-cut rules, but needs reference to a larger base of concepts so that a larger number of side assumptions (both in the inferential base and in the collateral commitments) become meaning-constitutive, in line with the theory-theory of concepts. Notice that the two views need not be in opposition (Laurence and Margolis, 1999, §7): different theories may in fact illuminate different aspects of conceptual content, the prototype theory explaining quick categorisations and typicality judgments, the theory theory explaining more considerate inferences and reasoning. With these caveats in mind, let us proceed.

What has inferentialism to say about the meaning of 'causes' *simpliciter*? Is there something that *all* causal claims have in common, irrespective of their relata and contexts of application? If there is, this is probably something thinner than envisaged by traditional monistic accounts. To the plurality of contexts and relata corresponds a plurality of low-level meanings of 'causes'. Still, isn't there something to be said about the *high*-level meaning of 'causes'? In chapter 6, I argued for the existence of one, vague notion of causality, but left open the issue of identifying this one, vague meaning. It is now time to address this issue by using the tools of IS. As I am going to show, there *are* some uncontroversial features, or 'typicalities' about the use of 'causes' to which everybody is committed. These typicalities help safely place some predicates outside the cluster and others closer to the centre of the cluster.

Recall that sentences have identical meaning iff one can be substituted for the other *salva inferentia*. The same substitutional principle applies to the case of subsentential locutions: "two subsentential expressions (...) share a semantic content just in case substituting one for the other preserves the pragmatic potential of the sentences in which they occur" (Brandom, 2000, p. 130). Predicates, in particular, are characterised by the following two features: syntactically, they are 'substitution-structural frames'; semantically, they have 'asymmetric substitution-inferential significance'. Let me explain.

Syntactically, the expressions '$p \to r$' and '$q \to r$' are substitutional variants of each other, that is, they belong to the same substitutional-structural *frame* '$\alpha \to r$'. Predicates are the particular substitutional sentence frames formed when singular terms are substituted in them. For instance, 'a causes b' and 'x causes y' are two substitutional variants of the same sentence frame 'α causes β'. *Semantically*, two expressions have the same meaning if one can be substituted for another, whilst preserving the status of all the material inferences in which they appear. That is, their frames have the same inferential significance. Singular terms are such that they can share the same meaning across all contexts. When this is the case (i.e., they stand for the same thing), substitution of one for another yields reversible material inferences. For instance, 'Benjamin Frankin' and 'the first postmaster general of the United States' have the same inferential significance. Both the inference from 'Benjamin Franklin invented bifocals' to 'the first postmaster general of the United States invented bifocals' and its reverse are materially correct. The thing is different with predicates: the inference from 'Benjamin Franklin walked' to 'Benjamin Franklin moved' is a good one, but its reverse is not. The inferential significance of predicate frames is *asymmetric*:

> That is to say that some predicates are simply inferentially weaker than others, in the sense that everything that follows from the applicability of the weaker one follows also from the applicability of the stronger one but not vice versa (Brandom, 2000, p. 135).

For instance, the circumstances of appropriate application of 'α walks' form a proper subset of those of 'α moves'. Now, let us indicate the meaning of 'causes' *simpliciter* with "$C_{<\alpha,\beta>}$", where α and β stand for any two singular terms. How do the above considerations apply to the analysis of '$C_{<\alpha,\beta>}$'?

Like other predicates, $C_{<\alpha,\beta>}$ is a substitution-structural frame with its own particular asymmetric substitution-inferential significance. Syntactically,

it corresponds to the frame 'α causes β'. Semantically, such a frame is (among other things) asymmetric. As per other predicates, there are two inferential asymmetries that are constitutive of its meaning, namely its downstream potential, or base-to-frame inferences (BF) and its upstream potential, or frame-to-target potential (FT).

Let us consider BF first. Let us denote with "$X_{i,<\alpha,\beta>}$" ($i = 1, \ldots, u$) the (binary) predicates that stand for test criteria, viz. those predicates from whose correct application 'causes' typically follows: 'regularly followed by', 'counterfactually depends on', 'raises the probability of', etc.[113] What is the relation that $C_{<\alpha,\beta>}$ bears to the criteria $X_{i,<\alpha,\beta>}$? According to the excess content thesis (§6.3), although there is no $X_{i,<\alpha,\beta>}$ such that $X_{i,<\alpha,\beta>}$ and $C_{<\alpha,\beta>}$ are incompatibility-equivalent, satisfaction of $X_{i,<\alpha,\beta>}$ does typically entail $C_{<\alpha,\beta>}$. Let us indicate the sentence frame '$< \alpha, \beta >$ satisfies some $X_{i,<\alpha,\beta>}$ or other' with the disjunction "$\vee X_i(\alpha, \beta)$". Then, $\vee X_i(\alpha, \beta) \in c^{\downarrow}$.

Let us now consider FT. The target of all causal claims comprises explanations, predictions and interventions *and* other claims (whose content varies from one causal claim to another). So, given 'A', a particular 'A causes B' may entitle one to 'A is to the right of B' as well as to 'B'. However, whereas all causal claims entitle to explanations, predictions and interventions, not all causal claims entitle to 'to the right' claims, so only the former belong to the target of $C_{<\alpha,\beta>}$. Let us indicate the sentence frames 'α obtains' and 'β obtains' with, respectively, "$O(\alpha)$" and "$O(\beta)$", and let us index α and β to times θ and τ, such that θ is prior to τ. Then, $< O(\alpha_\theta), O(\beta_\tau) > \in c^{\uparrow}$.

If BF and FT are put together, '$C_{<\alpha,\beta>}$' (that is, 'causes' *simpliciter*, in short SIM) can be—minimally—characterised as follows:

[SIM] $C_{<\alpha,\beta>} = \{< \vee X_i(\alpha, \beta), O(\alpha_\theta), O(\beta_\tau) > \mid \vee X_i(\alpha, \beta) \in c^{\downarrow}, < O(\alpha_\theta), O(\beta_\tau) > \in c^{\uparrow}\}$

That is, if one were committed to some test condition or other being applicable to some unspecific couple of relata $< \alpha, \beta >$, then one would typically be committed to 'causes' being applicable to $< \alpha, \beta >$; and if one were committed to c and the obtaining of one relatum, then one would typically be

[113] Notice the shift from talking of *sentences* in X to talking of *criteria* in X. The possibility of distinguishing between 'sentences in X' and 'criteria in X' was indicated in §6.7.1 as a *desideratum* of a plausible account. In the inferentialist account, this distinction is performed by the same procedure that is here used to distinguish '$C_{<A,B>}$' from '$C_{<\alpha,\beta>}$'.

committed (although to different extents) to the obtaining of the other relatum. Consider the claim 'Throwing the ball causes the window shattering'. 'Causes' in this claim is very close to SIM. Several test criteria are typically satisfied. That is, balls, throws and windows are typically such that throws are quite regularly followed by shatterings, shatterings are probabilistically and counterfactually dependent on throws, changes in throws' strength or direction result in changes in shatterings, shatterings depend on the arrangement of balls. And several target criteria are typically satisfied. That is, throws (resp. shatterings) often suffice to correctly infer shatterings (resp. throws) given some minimal knowledge of the context. As a result, the claim works as an exemplar, or attractor for the use of other claims. Its 'typicality' is measured in terms of the number of contexts (or the simplicity of the context descriptions) that make the material inferences correct. The larger the number of contexts, the closer the claim to the centre of the cluster.

Notice the difference between SIM and INF. First, SIM's base includes a disjunction of all test criteria. The more disjuncts are satisfied, the closer some token $C_{<A,B>}$ is to the centre of the cluster. This amounts to saying that at the centre of the causal cluster is a 'prototypical' notion of cause, which is correctly applied only if *all* test criteria are satisfied; further away from this prototype are other instances of causal predicates, which need not meet this requirement, and count as causal to a lesser extent—they still belong to the cluster but lie somewhere between the centre and the periphery. Notice the difference with DC (§6.3): in SIM, too, the more disjuncts are satisfied, the larger the causal content; however, causal content is not reduced to such disjuncts. Secondly, SIM's target includes the consequences that all causal claims have in common, namely the inference to the obtaining of one relatum granted by the claim plus the obtaining of the other relatum. This is what keeps together the various $C_{<A,B>}$. In particular, SIM does not specify what Y collateral commitments must be among the premisses that grant FT inferences. The reason is that the appropriate Y may change from case to case. The mastery of $C_{<\alpha,\beta>}$ only presupposes the ability to infer the effect given the cause and vice versa. SIM rules that the prototypical notion at the centre of the cluster is such that $c^{\uparrow}$ contains collateral commitments which allow both *both* forward- and backward-looking inferences, whilst guaranteeing (among other things) the inferential asymmetry of $C_{<\alpha,\beta>}$. Other instances of 'causes' and other predicates will be more or less distant from this prototype depending on how much their upstream potential differs from the upstream

potential of the prototype. Let me now discuss the virtues, if any, of SIM. If SIM is to provide an adequate criterion for a predicate to count as causal, it should be both *flexible* and *informative*.

First, the criterion should be flexible enough as to accommodate the variety of low-level, context-specific claims with an obvious causal significance that one encounters in everyday as well as scientific parlance. In particular, scientists are often reluctant to use explicit causal vocabulary, and they rather talk of 'promoting', 'activating', 'inhibiting', etc. For instance, as regards the result of their study, Legewie et al. (2006) state that 'XIAP-mediated feedback *cooperates* with Casp9 cleavage by Casp3 to *bring about* bistable and irreversible Casp3 activation' (*ibid.*, p. 1068, emphasis mine). SIM correctly rules that the claims describing such relations have causal content: for any $< \alpha, \beta >$, the predicates that describe the $< \alpha, \beta >$ relations would follow from X-contexts, and would commit to inferring β given α and vice versa. Closer to the centre of the causal cluster one typically finds predicates that meet the above constraints in a number as large as possible of contexts. Notice that a lot of predicates will turn out causal according to SIM. But this is nothing to worry about. On the contrary, it confirms why an inferentialist analysis of 'causes' is appropriate: the non-causal vocabulary which could be used to give a reductive analysis of 'causes' may be too limited. Indeed, the 'sparse base' argument is one of the main motivations for an anti-reductionist position about causality (see Carroll, 2009, pp. 285-286).

Secondly, at the same time the criterion should provide an informative analysis of 'causes' *simpliciter*, viz. of the high-level meaning of 'causes'. In order to do this, the criterion shouldn't be *too* flexible. This means, among other things, that it should say something on what 'causes' does *not* mean, so that certain predicates are ruled out as non-causal. The inferences in SIM involve relata that stand in an asymmetric relation, fixed by indexing them at different times. Since semantic, logical and mathematical predicates apply to same-time relata, they automatically turn out non-causal. Predicates describing empirical relations with little, or no causal content, such as 'to the right of' or 'higher than' are non-causal because they would not typically follow from the correct applicability of X predicates. And then there are in-between cases, viz. predicates whose meaning overlaps with 'causes' but also differs from it in important respects. One example are predicates that indicate constitutional relations, e.g., 'constitutes', as applied to the pair <two H atoms and one O atom plus their structural relations, an H_2O molecule>. For

relata obtaining at the same time, 'constitutes' is usually non-causal. An exception may be homeostatic mechanisms where the micro-level configuration continually sustains the macro-level equilibrium, e.g., a limit cycle. In such cases, it seems appropriate to say the the former causes the latter—but in a way that need not involve any vicious circularity (§2.2.1). Instead, for relata obtaining at different times, so that the composition at one time is responsible for the whole at some other time, 'constitutes' is typically causal. In complex systems, the asymmetries between the description of microstates and of macrostates are often crucial for predicting or explaining their behaviour. In all such cases, 'constitutes' has an obvious causal content. A final class of in-between cases are the predicates that correspond to traditional, monocriterial analyses of the meaning of 'causes', such as 'is regularly followed by' or 'raises the probability of'. Since in this case the target predicate X_i also appears in $c^{\downarrow}$ of SIM, one must rely on (lack of) incompatibility equivalence between 'causes' and X_i to pinpoint the differences in meaning, along the lines of 7.1.2. Instead, to show the similarities in meaning one can remove X_i from $c^{\downarrow}$. Then, the applicability of the remaining predicates does typically entail the applicability of X_i in the same circumstances in which it does entail the applicability of 'causes'. And there are differences in FT inferences, too: the applicability of X_i does typically—but *not always*—entitle one to the same forward- and backward-looking inferences to which 'causes' entitles. This is enough to show that there is no perfect overlapping, hence no X_i provides an exhaustive analysis of 'causes' *simpliciter*.

Importantly, the informativeness of the inferentialist account depends not on the fulfillment of conditions as strict as those imposed by reductive accounts, but on the satisfaction of several criteria, which are weighted differently depending on the situation and on the claim's coherence with our background commitments. This makes of SIM an instance of 'cluster-concept account' of causality. For Woodward, in such accounts 'causes' is "vague" and its application "contestable" (Woodward, 2003, p. 91). Although I concede this much to Woodward, I deny that this prevents one from telling an illuminating story on why the criteria in the cluster are grouped together and we have *one* notion of cause. The inferentialist story goes like this: 'causes' is one—although vague—concept in virtue of its role of licensing inferences about predictions, explanations and interventions, and test criteria which make 'causes' explicit are grouped together because they all contribute to licensing such inferences. That the story is informative is illustrated by com-

paring SIM and DC. Whereas DC accounts for counterexamples by adding disjuncts to the analysis ex post, but without explaining what makes them genuine instances of the concept, SIM explains their belonging to the cluster in terms of similarity to base and/or target criteria. For instance, SIM may be used to justify that relations which involve neither difference-making nor production may still count as (borderline) instances of causation provided they, e.g., satisfy all target criteria. So, SIM proves more explanatory and less ad hoc than DC.

7.3 The secret (?) connexion

A worry may be that, if one retreats to analysing causation in terms of inferences drawn by language users, one stops asking the important question 'What is causation?' and contents oneself with the more modest—and philosophically less relevant—question 'How do we use the notion of causation?' In other words, one gives up on the attempt to identify the secret connection that underpins the causal relation.

Let us step back for a moment. Recall that it is the search of the secret connection which has led to many of the analyses rebutted in chapters 4–6. An inferentialist may observe that the *symptom* of the problem is that all parties for which the secret connection issue is relevant tie the meaning of causal claims to the possibility to *experience*, or *define*, the necessity of the connection, and that the skepticism on causality is the result of failed attempts to identify such a necessity. Here is an inferentialist *diagnosis*: the issue arises from a bad intuition on what counts as meaning of something and the—implicit—adoption of one or the other non-inferentialist semantics as applied to 'causes'. Needless to say, the inferentialist rejects both non-inferentialist semantics and the intuitions that motivate them.

One such semantics is *semantic empiricism*. This is, roughly speaking, a theory of meaning based on the Humean distinction between concepts that derive from 'impressions' and those that don't (e.g., 'causal connection'): only to the former category there correspond existents. The implicit adoption of semantic empiricism may explain certain attempts—both realist and anti-realist—to reduce causal talk to a basic 'impression' of force, or pressure. Now, one may grant that it is legitimate to infer some causal relations on the basis of impressions. The point is that the low-level causal relations we can have an impression of (e.g., your push causing me to move) do not seem to

be the same causal relations science talks about (e.g., p53 causing apoptosis, chartist behaviour causing crashes). No wonder then that it is hard to find out the secret connection in the latter case. And no wonder that attempts to explain causation as involved in higher-level causal laws in terms of basic, low-level causal relations leaves the secret connection mysterious. But this, for the inferentialist, doesn't show anything on the nature of the 'secret' connection, only illustrates the inadequacy of representationalism.[114]

A similar consideration applies to another theory of meaning which seems implicitly presupposed by much of the philosophical debate on the metaphysics on causality, namely *verificationism*. For the verificationist, the meaning of a statement is reducible to the set of observations which either verify or falsify it. Corollary: if causal claims are to be meaningful, then they must have necessary and sufficient verification conditions. However, it is well known that the verificationist criterion has been abandoned as a meaningfulness criterion after realising that claims containing theoretical terms are not reducible to sets of observations which are necessary and sufficient to verify or falsify them—there are no such sets. So, why is verificationism still implicitly presupposed by philosophical explications of 'causality', which is arguably further away from observation than theoretical concepts?[115]

This old-style verificationism is nowadays replaced by the view that meaningful claims have (at least) necessary and sufficient assertibility conditions (Dummett, 1991), e.g., canonical ways of being introduced (verification conditions) and eliminated (use conditions). Although meanings of expressions are not typically reducible to sets of observations but are fixed by reference to same-order expressions, meanings stand in a relation of partial ordering, such that any non-canonical way of verifying an expression contributes to meaning only insofar as it is reducible to a canonical one. If this partial ordering is possible, one may still neatly distinguish between meaning-constitutive inferences and non-meaning-constitutive inferences. Neo-verificationism is subject to the following criticism. Borderline concept instances constitute 'anomalies' with respect to the conceptual core fixed by the canonical rules. Their content cannot be explicated by reference to canonical rules only. This hampers the possibility of neatly distinguishing meaning-constitutive (canonical) uses from non-meaning-constitutive (non-canonical) one.[116]

[114] For Sellars' critique of the Humean theory of meaning, see Sellars (1962, pp. 50-ff.).
[115] Reiss (2009b) offers a very similar argument to the same point.
[116] For Brandom's own concerns about verificationism, see Brandom (2000, pp. 63-66).

In the light of this diagnosis, the inferentialist can also recommend a *cure*, based on the rejection of semantics which reduce meaning to representational relations or verification conditions. When inferentialism is embraced, the secret connection need not be 'secret' anymore, and can be made explicit as any other part of language. What is special about causality is that its meaning must be made explicit via *more inferences* and in a *less intuitive* way than in the case of other concepts. The cure can be accompanied by some argument to the point that (causal) language has a referential function after all. Only reference *follows* from the theory of meaning rather than *grounding* it (see 8.1.2). Once we realise that there is no secrecy in the connection, the puzzlement deriving from the attempt to understand what makes the connection *necessary* disappears, too. For the inferentialist, the necessity has to do with the normativity of the inferences. *This* sort of necessity is—for those who embrace analytic pragmatism—easier to grasp and accept.

Conclusion

The task of providing a satisfying account of causality is very challenging. In this chapter, I have shown that an inferentialist account has the resources to face the challenge. It can make room for the contribution of both test conditions and use conditions to the meaning of causal claims. It can be flexible enough as to accommodate the variety of nuances that causal talk takes on in different areas of inquiry, thereby accommodating the idea that causality is, in a sense, multi-faceted. And it can be informative enough as to point to the core features that all causal claims have in common, thereby doing justice to the idea that, in another sense, causality is one, although vague, concept. In the next chapter, I apply this framework to the analysis of causality in complex systems.

Chapter 8

Causality in Complex Systems

In this chapter, I discuss the *meaning* and the *objectivity* of causal claims in complex systems sciences. First, I consider how the inferentialist can talk of causal relations as being more or less objective. I do this by reinterpreting the objectivity of the causal relation in terms of the correct assertibility of the corresponding causal claim (§8.1). Then, I turn to discussing the meaning and the objectivity of causal claims in systems biology (§8.2) and computational economics (§8.3). Finally, I discuss whether the inferentialist notion of objectivity as based on normativity is suitable to account for the objectivity of causal claims presupposed in everyday and scientific discourse (§8.4).

8.1 The objectivity of causal relations

8.1.1 Objectivity as correct assertibility

How should one explain one's judgement that certain causal claims are—to some extent, or to a good extent—*true*? And how should one explain one's judgement that certain models get the causal story *right*, or correctly *represent* the mechanism responsible for the behaviour?

For the representationalist, the meaning of an expression, e.g. 'causes', is its contribution to the truth conditions of the sentences in which the expression appears. Correspondingly, what is signified by the expression, e.g. the causal relation, is objective to the extent that it contributes to the state of affairs that makes the sentence true. In other words, causal relations are objective if causal claims represent the world as it is, i.e., is, if 'causes' is referential and causal facts belong to the ontological furniture of the world. Here 'objective' has an *ontological* meaning, viz. it concerns a mode of existence of a class of entities or the world as a whole.

In contrast, for the inferentialist reference is not what grounds objectiv-

ity, but rather a consequence of it.[117] Objectivity, in turn, is interpreted as having to do not with truth-making relations between claims and states of affairs,[118] but with *correct assertibility*, that is, correct use of the claims as premisses or conclusions of arguments. More precisely, 'objective' has a *semantic* meaning, concerning the basis for distinguishing between what *seems* to be correct and what *is* correct in the application of assertive content (see Skovgaard Olsen, 2012, §1). The correctness of the assertions cannot be adjudicated from *outside* the space of reasons, but involves a self-correcting process—both epistemically and semantically constrained—*within* the space of reasons. The epistemic constraints (priority of observational knowledge, results of actions) allow that *any* claim may be put in jeopardy, only not all at once. The semantic constraints (the rules internal to the game, viz. compatibilities and incompatibilities between claims) fix what one ought to endorse or is entitled to endorse. In practice, this process tends to remove or alleviate disagreement and lead to 'harmony', which may be informally characterised as a sort of reflective equilibrium between introduction and elimination rules, which obtains when successful extra-linguistic navigation and lack of intra-linguistic disagreement on commitments and entitlements make the rules entrenched (cf. Brandom, 2000, pp. 72-76).

So, when it comes to causality, whether a causal relation is objective becomes a matter of semantic justification, a matter of having reasons for endorsing the corresponding causal claim in the arguments where it appears. This entails answering the question 'Is the causal relation objective?' by answering the question, 'Is the causal claim correctly assertible?' And since causal relations are context-sensitive, the more fine-grained questions to ask become: Is c correctly assertible in such-and-such a context? To what extent is the assertibility of c context-sensitive?

Since causal relations are contextual, no general answer can be given. The reasons that justify the assertion of one causal claim in one context differ not only from the reasons for asserting another causal claim, but also from the reasons for asserting the same causal claim in another context. This is because the justificatory relation between such reasons and the claim itself is a matter of *content* of what is asserted by, respectively, the reasons and the

[117] For more on the consequences of the inferentialist account with regard to the representational force of causal claims and models, see §8.1.2.

[118] Among traditional criteria for objectivity are not only truth conditions, but also epistemically constrained notions of truth, such as belief in the long-run (Peirce), belief that remains satisfactory (James) or credible (Goodman), rational belief (Putnam), etc. (see Skovgaard Olsen (2012, §1) and references therein).

claim. More specifically, the goodness of the inferences from the evidence to the causal claim, and from the claim (together with the collateral commitments) to its possible consequences, depend on *material* conditions, not just formal conditions. That is, such inferences do not obey any ready-made logic by which their validity can be 'calculated' by simple consideration of the sentences that stand in the entailment relation. So, one can't formulate fully general assertibility conditions that do not depend on the specific relata and context of application. And this is not a special fact about causal claims, but a general fact about all non-logical claims. Ultimately, one is always left with a *decision* on whether or not to endorse the claim on a given occasion. But this does not mean that the conditions for correct assertibility are arbitrary. The existence of objective conditions of assertibility presupposes a space for the rational evaluation of the decision to endorse or not to endorse as more or less justified. More precisely, it presupposes having the resources to *vindicate* such a decision in some *principled* way, an issue to which I now turn.

Understanding what having such resources amounts to requires drawing a distinction between a *first-person* perspective and a *social* perspective on correct assertibility, and the different tools available to the speakers from the two perspectives. On the first-person perspective, the objective assertibility of c depends on commitment to c (given available evidence) and entitlement to c (given collateral commitments). Obviously, whether or not one takes c to be assertible, one may be mistaken. The grounds for the evaluation of the correctness of the assertion are not private, but public. The evaluation takes place in the social space that commitments and entitlements institute (Brandom, 2000, chap. 5). To use an analogy with games, this evaluation works as follows: a participant in the game makes a move; if other participants in the game disagree with his move, they can challenge him and ask for reasons; it is then up to him to provide such reasons, possibly reasons such that the other participants, too, are committed to them, so that they cannot deny the correctness of his move. During the game, the players keep track of each other's 'score', viz. the commitments and entitlements that their use of the language presupposes. The possibility for them to do so, in spite of their different sets of beliefs, rests on the possibility to discuss the propositional content of the assertions by using two different modes of ascription, viz. *de dicto* and *de re*.

De dicto ascriptions are introduced by 'that' clauses. They makes explicit what part of the commitment is attributable to the attributee. For instance,

Andy can ascribe to Bob the belief that the Earth is flat, whilst not believing it himself. Andy can express this by saying 'Bob claims *that* the Earth is flat'. *De re* ascriptions, instead, are introduced by 'of' clauses. They make explicit the part of the commitment that is adopted by the scorekeeper. Bob may or may not believe that if one were to sail West one would not return to where he started his journey. However, this is what his belief that the Earth is flat commits him to. Andy can express this by saying 'Bob believes *of* the Earth that it is so shaped that if one were to sail West one would not return to his starting point'.

The *de dicto*/*de re* distinction is what grounds *reference*, which is something that language permits by providing us with the resources to talk *of* things from within a social perspective. It is in virtue of this social dimension that, for Brandom, propositional content is *necessarily* representational content. This is because communication presupposes the possibility to agree or disagree on the objects of our beliefs on the basis of sets of commitments attributed or undertaken (see Brandom, 2000, p. 183).

For instance, both Andy and Bob believe that the Earth is a planet. Although they have different theories, or 'conceptions', of the Earth, they can discuss about the same thing, ascribing to each other beliefs either *de dicto* or *de re*. The ascriber will use the *de dicto* mode if he has positive reasons not to believe, or has no reason to believe—and wants to remain neutral on—the content of the proposition he is ascribing. He will use the *de re* mode if he wants to emphasise his commitment to something, whether he agrees on what he ascribes (e.g., that the Earth is a planet) or not (e.g., that the Earth is flat).

Although playing the language game presupposes that the speakers are committed to the idea that there is a fact of the matter as to whether their assertions are true or not, and what the Earth is like, at no point of the game truth assessment need involve the attribution of a truth property. What goes on, for the inferentialist, is continuous scorekeeping, giving and asking for reasons, articulating distinctions between what is correct and what seems to be correct.[119] In Brandom's own words:

> The practical navigational capacities that are made explicit in *de re* specifications of the contents of ascribed propositional commitments express the standing commitment each of us has to

[119]The reader interested in the details of Brandom's machinery is referred to (Brandom, 1994a, chap. 8).

> their being *one* set of inferential roles that bind *all* interlocutors: those, namely, determined by multipremise inferences in which the collateral commitments supplying auxiliary hypotheses are *true* (Brandom, 2007, p. 670).

It is in this sense that speakers with different concepts of 'causes' or ascribing different meanings to the same causal claim are nonetheless bound to the same rules, or criteria, for the assessment of their correct applicability. I give a characterisation of the criteria invoked in the assessment of the objectivity of causal relations in §8.1.3. I introduce this topic by describing in §8.1.2 how inferentialism can be usefully applied to account for the objectivity of *scientific* claims—of which causal claims constitute an important instance.

8.1.2 The (inferential) rules of science

Scientific reasoning—more precisely, the rules that regulate the acceptance of scientific hypotheses—can be interpreted along inferentialist lines. One attempt to do so is in (Zamora Bonilla, 2006). Here, scientific rules are classified as 'argumentation rules' (viz. language-to-language rules), 'entry rules' and 'exit rules'. They have the function to introduce, evaluate, modify and put to use the 'deontic scores' of scientific claims. Deontic scores measure the recognition that the scientific claim receives from the community. Zamora Bonilla interprets this recognition as the aggregation of two components: an 'internal' score, that measures how the claim follows from the inferential rules and previous commitments of the community; and an 'external' score, that measures how the claim coheres with the community's inferential practice.[120] Argumentation rules change a deontic score into another (e.g., add or remove a commitment). Entry rules determine how deontic scores are affected by events (e.g., evidence gathering, authority reports). Exit rules determine how the claims included in the deontic score transform into obligations to perform or abstain from performing certain actions (e.g., allocation of resources, experiments). Together, these rules generate a loop, comprising perceptions,

[120] Zamora Bonilla suggests that number of favourable citations and of published papers be used as indicators of, respectively, internal and external scores (see fn. 10). It should be noted, however, that although the distinction is in principle useful, in practice there may not be a sharp divide between the reasons that bear on internal and external scores. For instance, a citation, where the cited claim appears as a conclusion reached by other means, may contribute to the internal score, too. And a publication, such that the claim is particularly coherent with the referee's commitments, contributes to the external score, too. So, it may well be that the same reason can affect the two scores at once.

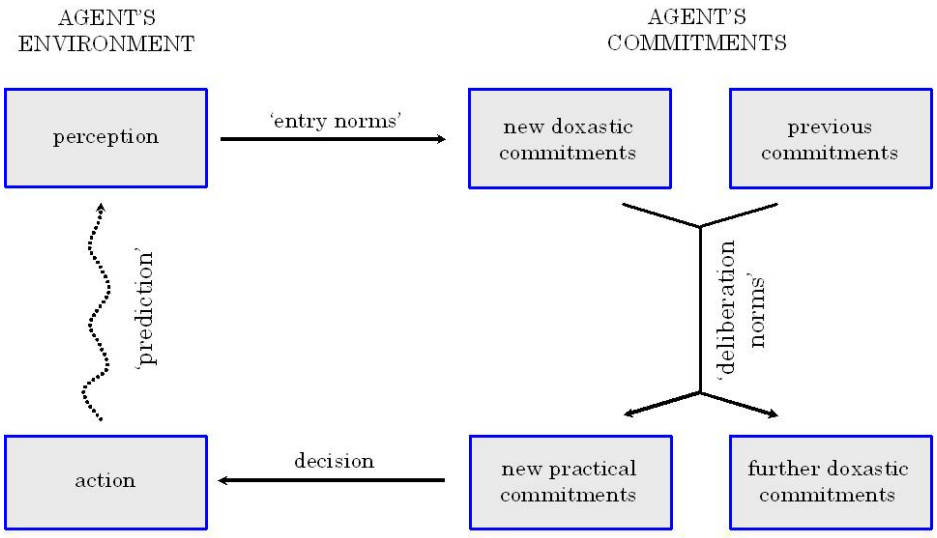

Figure 8.1: A normative inferentialist view of rational action. Redrawn from (de Donato Rodríguez and Zamora Bonilla, 2009a, p. 105) (cf. figure 2, Zamora Bonilla, 2006, p. 194).

the resulting doxastic and practical commitments and the outcomes of actions, which determines the acceptability of the claims (figure 8.1).

This framework allows the inferentialist to account for the representational force of scientific claims, including the *causal* ones. This force, for a community of scientists helping themselves to the *concepts* of 'causal relation', 'complex system', 'mechanism', etc., is constituted by the *conceptions*, or theories, corresponding to such concepts. Such conceptions may vary to some extent from one scientist to another, and from one community to another. Also, these conceptions may be incompletely or incorrectly specified. Incompletely, because only part of this meaning is actually made explicit by way of models and relations depicted therein. Incorrectly, because certain parts of this meaning may depend on incompatible commitments, or less-than-optimal (both extra- and intra-linguistically) commitments. Either way, the conceptions that make explicit the meaning of causal claims may not include all meaning-constitutive inferences. Still, there is a sense in which such inferences are part of the concepts, insofar as the speakers are committed to them by their use of the language, whether or not they are aware of them.

With this framework in place, one can give a (hopefully more illuminating) answer to the question: Does 'causes' refer? As said, for the inferentialist

reference of 'x' is grounded in the possibility to communicate about x in *de re* mode, e.g. 'S believes *of* x that $\phi(x)$' in a way that promotes 'harmony'. So, we may say that a concept 'x' refers if its 'conception', or theory ϕ, is *harmonious*. In this respect, 'causes' is a peculiar case (cf. §6.7.2). On the one hand, the relative stability of kinds of purposes (Z) unify ϕ('causes') in a way that makes of causality a pivotal concept in our conceptual apparatus. Causal claims are often successful licenses to predictions, explanations and interventions. Agreement on typical exemplars of causal relations promotes harmony, and make 'causes' *referential*. On the other hand, the diversity of contexts of application (X and Y), on which the existence of more or less numerous or successful consequences seem to depend, disunify ϕ('causes'). The resulting difficulty in making explicit the boundaries between causal and non-causal facts makes harmony difficult, and 'causes' only *vaguely* referential.

A nice case can be made for an inferentialist understanding of the representational force of models of complex systems. Recall Rosen's *modelling relation*, mentioned in chapter 1.3.4. The complex systems scientist subscribing to Rosen's picture envisages modelling as a complex practice, "the art of bringing entailment structures into *congruence*. That is, the formal description, encoding, implication and decoding must be congruent withe the causal events being modeled in the real world" (Mikulecky, 2001, p. 346). However, the complex systems scientist understands this 'bringing into congruence' not in terms of some naïve one-to-one correspondence relation taking place outside the space of reasons, but in context, from within the perspective of the scientist who builds and uses the model. A model is interpreted as a formal surrogate of portions and aspects of reality with the essential function of aiding the modeller's reasoning. It acquires meaning only in the context in which it is interpreted and used. So, if one wants to understand the relation between models and systems and what the models can tell about the causal relations in their targets, one must have a prior understanding of the modelling relation, as a complex relation between targets, formal structures *and modellers*.

This complex, intrinsically context-dependent entanglement among model, system and the modeller's activity is well captured by an inferentialist account of scientific representation (Suárez, 2004; de Donato Rodríguez and Zamora Bonilla, 2009a). This characterises scientific models as tools for surrogative reasoning (figure 8.2). First some features of interest are selected from the target system and translated into a formal structure, which can be

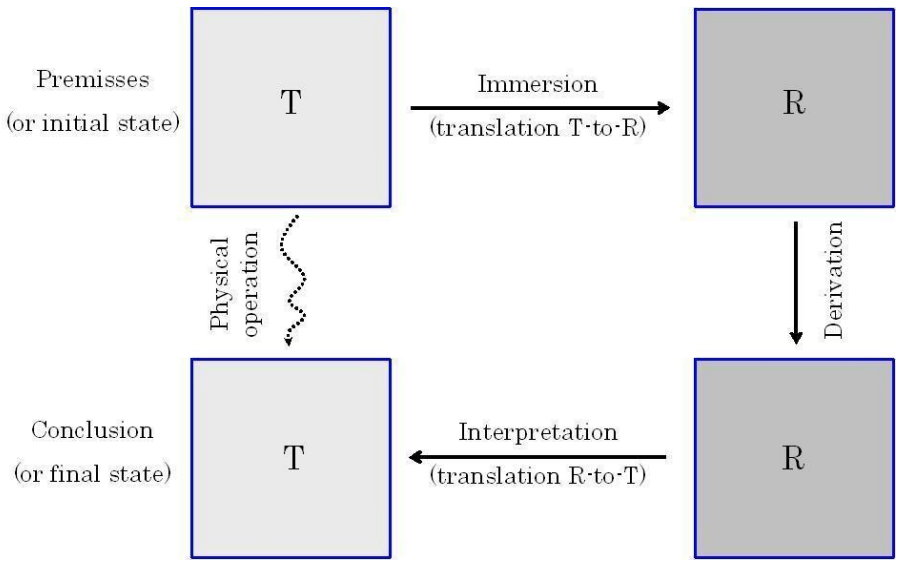

Figure 8.2: The three steps in the making of inferences with the aid of models. T = target; R = representation. Redrawn from (de Donato Rodríguez and Zamora Bonilla, 2009a, p. 103).

anything from a mental model to a map to a mathematical model. This step, called 'immersion', corresponds to what complex systems scientists call 'encoding'. Then, the behaviour of the formal system is derived. Finally, the derived behaviour is translated back onto the target system to check whether the derivation faithfully mimics the physical operation that goes on in the system. This final, third step is called 'interpretation', and corresponds to the complex systems scientist's 'decoding'. Importantly, de Donato Rodríguez and Zamora Bonilla (2009a, p. 103) explicitly state that the inferences drawn within the model often aim to reproduce the *causal connections* between the real events occurring in the target system—although their model is general enough as to include systems, such as maps, in which it is not the causal structure that the model aims to reproduce.

Now, in what sense is this an *inferentialist* account of representation? Why can't one say that all there is to the meaning of the model is its capacity to map successfully onto the system's structure, a fact which is only incidentally dependent on whether or not the scientists happen to stumble on the right model? In other words, why should one regard the meaning of the model as a three-place relation between model, target and modeller, rather than a two-place relation between model and target only? The reason, for the

inferentialist, is that if one fails to consider the role that the model plays in the life of those who produce and use it, one misses a crucial aspect of what confers to it its representational force. To account for what gives the model such a force, the inferentialist offers a more complex story that tells (in the metalanguage) how the model helps the agents relate to their environment, according to the rules described above and depicted in figure 8.1.

On this account, commitments and norms of agents and communities evolve. As a consequence, also the meaning of the claims caught up in the modelling loop evolves. 'Good' scientific models will be those that lead to 'satisfactory' results, or better results than other models: first, the model should increase the ratio of successful inferences; secondly, it should increase the number and variety of inferences we were able to draw from the rest of our commitments; thirdly, it should reduce the cognitive or computational costs of the activity of drawing consequences. Does the model refer? The degree of 'realism' of the model is explained in terms of the ranges of circumstances where the representans can be substituted for the representandum to draw inferences by means of the model rather than a black box. So, the reality of a complex system is weaker than that of, say, a chair. The contextuality of complex systems, that is, the sensitivity of their identity and their behaviour to the context, is accounted for and made explicit in terms of features of the meaning (as inferential role) of the models by which we refer to them.

I now turn to characterise in more detail the 'rules' by which *causal claims* are judged satisfactory (read: correctly assertible). The reader is invited to interpret such rules as a characterisation of the reasons that bear on the aforementioned deontic scores in the case of causal claims as explicated in §7.2.3.

8.1.3 Characterising objectivity

In what follows, I will describe general criteria that underlie the correct assertibility of causal claims and characterise the objectivity of the causal relations described therein. Correct assertibility depends on *two* normative statuses and their interactions, namely commitment and entitlement:

1. Ought one be committed to c given X? To what extent do changes in X affect entitlement to c?

2. Is one entitled to z given Y and c? To what extent do changes in Y or c affect entitlement to z?

Since there are two normative statuses on which correct assertibility depends, there are also two dimensions along which correct assertibility should be characterised, namely *commitment* and *entitlement* to certain conclusions given certain premisses. Along both dimensions one should consider what makes for the plausibility of the material inferences, that is, the reasons that make an assertion put forward as *prima facie* correct, or as *seemingly* correct, actually correct, or justified.

Commitments first: What commitments grant commitment to the conclusion? If c is correctly assertible, the argument in which it appears should ideally remain correct if there is some change in the premisses (i.e., substitutions, enlargements, shrinkages) and c is kept fixed, whether as a premiss or as a conclusion. The intuition behind this criterion is that if an argument is success-conducive if certain premisses are included, and whether or not certain other premisses are, then one will assume that the former not the latter make the argument correct in the circumstances in which this has a bearing.

Let us first consider the case where c appears as a conclusion, viz. the downstream potential of the causal relation. Here c should robustly follow across changes in the premisses. One can interpret this as a requirement that c follow from independent tests, or tests performed on different populations, in different contexts, etc.:

$$X_1 \vDash_{\text{Inc}} c \tag{8.1}$$

$$X_2 \vDash_{\text{Inc}} c \tag{8.2}$$

$$X_3 \vDash_{\text{Inc}} c \tag{8.3}$$

$$\ldots \tag{8.4}$$

So, 'smoking causes cancer' should belong to the set of one's commitments if the association is significant among humans, irrespective of gender, age, diet, ethnicity, etc., if the relation is confirmed by animal studies over different species, etc. Having the claim among one's commitments allows one to make sense of a variety of observations, and makes a good candidate inference ticket, which can then be spent in a variety of circumstances.

The other case is where c is one of the premisses, viz. the upstream potential of the relation. Here z should robustly follow from c across changes in

the set of collateral commitments Y, despite the non-monotonicity of $\models_{\text{Inc}}$:

$$c, Y \models_{\text{Inc}} z \tag{8.5}$$
$$c, Y, a \models_{\text{Inc}} z \tag{8.6}$$
$$c, Y, b \models_{\text{Inc}} z \tag{8.7}$$
$$\ldots \tag{8.8}$$

So, if the claim (e.g., 'smoking causes cancer'), together with the occurrence of the cause as well as background knowledge, allows one to infer a given causal effect (e.g., incidence rate of lung cancer), and the inference remains success-conducive even when the influence is ignored of other, potentially causally relevant factors, or confounders, present in the situation, we have reasons to believe that the success of the inference depends on the claim, which is an inference ticket worth having.

Let me now turn to the entitlements: What lack of commitments removes entitlement to the conclusion? If c is correctly assertible, the arguments should 'ideally' turn incorrect if the other premisses are kept fixed but c is removed from the set of the premisses:

$$c, Y_1 \models_{\text{Inc}} z \tag{8.9}$$
$$c, Y_2 \models_{\text{Inc}} z \tag{8.10}$$
$$\ldots \tag{8.11}$$

but

$$Y_1 \not\models_{\text{Inc}} z \tag{8.12}$$
$$Y_2 \not\models_{\text{Inc}} z \tag{8.13}$$
$$\ldots \tag{8.14}$$

This criterion demands that c be included in the premisses if one wants to be entitled to z given Y. In other words, given Y, assuming c is necessary to infer z. So, for instance, it would not be possible to design effective policies on smoking habits to decrease the incidence of lung cancer in a given population, if it were not for the knowledge that smoking *causes* lung cancer, and in that population it does so by way of certain intermediary steps, contributory or inhibiting factors, etc.

In sum, objectivity can be characterised as *robustness* of the meaning-constitutive inferences.

8.1.4 Challenging objectivity

Since there are several ways in which a causal relation may be objective, there also are several ways in which such an objectivity may be challenged, namely by challenging the role of c as a premiss or conclusion in the arguments in which c appears.

The case to which most attention has been devoted in the philosophical literature is the case of the validity of the arguments where a conclusion as regards the obtaining of the causal relation is established. In this case, questioning the objectivity of the relation amounts to questioning one's grounds for taking the causal claim as warranted.

In science, this issue is traditionally associated with the internal validity of the arguments warranting scientific hypotheses, namely with the correctness of the methodology by which the hypotheses are tested (or supported, or confirmed). Consideration of this scientific fact has typically given rise to the more philosophical task of identifying the test conditions of the hypotheses being established with their truth conditions. Lurking in the background was the idea that the non-controversial conditions X used to establish a claim about the existence of some controversial object C (e.g.: theoretical entities such as electric charge; or causal relations; or...), could also be used to reduce the objectivity of the controversial object C to the objectivity of the non-controversial one (e.g., the voltage measured by a meter attached to a conductor; the probability raising of the effect by the cause). In this way, not only is the inference from X to C part of what it takes for C to be objective, but so is the inference from C to X.

However, scientific practice need not be so interpreted. On the contrary, it is more faithful to scientific practice to distinguish the kind of objectivity involved in issues of internal validity from the objectivity involved in issues of external validity, which concerns the exportability of the results obtained in test situations, by means of some method, to some target situation. In Cartwright's words, "the inferences that are licensed from that method are tied to the populations and situations in which the evidence is obtained and license to go beyond those must come from somewhere outside that method" (Cartwright, 2007b, pp. 38-39). So, internal and external validity obey dif-

ferent 'logics'. The only thing that they have in common is that a failure depends either on a lack of entitlement preventing another entitlement, or on a commitment removing an entitlement.

Challenges to the assertibility of c in claims of *internal* validity are of the following kinds:

- entitlement to c only if $x \in X$, but no entitlement to x;
- commitment to x, and no entitlement to c if $x \in X$.

Instead, challenges to the assertibility of c in claims of *external* validity take different forms:

- entitlement to z only if $y \in Y$, but no entitlement to y;
- commitment to y, and no entitlement to z if $y \in Y$.

In general, a failure of external validity is such that for some reason one of the facts that was observed in the context of application of c, and granted the correct applicability of c, fails to manifest in the context of using c. For instance, one may expect that, since probability raising granted c in some test situation, c will in turn entitle to infer to probability raising in some target situation. Yet, something goes wrong. Let us indicate with d the claim that one relatum makes a difference to the other. The issue of external validity is that c is *insufficient* to draw conclusions on d. More formally put:

$$X, d \vDash_{\text{Inc},K_{test}} c \tag{8.15}$$

but

$$c, Y \nvDash_{\text{Inc},K_{target}} d \tag{8.16}$$

There are two main reasons why external validity fails, and the inference to the target situation goes wrong. First, there is the difference between the test population and the target population. For instance, one can take the claim 'Smoking causes lung cancer' as well established on the basis of an RCT on mice, after observing an overall probability raising of cancer across the population that was assigned to treatment (injection of tar in the lungs, or the like) with respect to the control population. One may then ask whether the difference-making relation that was observed in the test situation will also be observed in some target population, following to, say, a policy that is meant to exploit the tested relation. For instance, one

may ask whether banning smoking in some target population of humans will result in a decreased incidence of cancer with respect to the cancer level that would have been observed in the same population had the policy not been enforced.[121] Here the inference may go wrong due to the *context sensitivity* of the relation, such that c contributes to the correctness of the inference only when $K_{test} \approx K_{target}$, not when $K_{test} \neq K_{target}$.

The second reason for the failure of external validity has to do with the *faithfulness* of the relation. In spite of $K_{test} \approx K_{target}$, the difference-making relation shows up in test situations (via patterns of regularity, or counterfactual dependence, or probabilistic difference, etc.) not in target situations, due to multiple or neutralising capacities, redundancies and back up mechanisms, etc. Suppose one has established that contraceptives prevent thrombosis via the prevention of pregnancy. And suppose that, besides this component effect, contraceptives have another component effect on thrombosis, namely the release of a harmful chemical in the bloodstream, which promotes thrombosis. Here, one can fail to observe a decrease in thrombosis among women who take contraceptives if the former component effect is neutralised by the latter, so that the net effect is null. Or suppose one has established that gene 1 causes a disease. And suppose that the operation of gene 1 normally trumps the operation of gene 2, which also has the capacity to bring about the disease if gene 1 does not operate. Here, trying to eliminate the onset of the disease by a knock out of gene 1 won't do, because gene 2 will act as a back up. Failures of faithfulness are such that $x \in X$ but $x \notin Z$.

A third case, still having to do with the upstream potential of a causal claim, and which has been largely neglected in the philosophical literature, involves the capacity of the causal claim to assist in inferences whose conclusion does not concern the obtaining of one or the other relatum, but some other state of affairs, e.g. another causal claim or claims on the obtaining of something which is not a relatum. One may view this as having to do with the external score of the claim, rather than its internal score. The more often c contributes to the correctness of such inferences, the more appropriate it seems to assign to C a stable place in our set of commitments. Correspondingly, the more often c fails to make a difference to other inferences, and to cohere with them (by being isolated, or 'disconnected', from them), the less useful it is to include C among our commitments. Here, the failure of

[121] This counterfactual value can be estimated by looking at past values in the same population, or by looking at values in populations which are as similar as possible to the target one, etc.

objectivity depends on c being unnecessary or insufficient for z given Y:

- unnecessary, if given Y, z follows irrespective of c

$$c, Y \vDash_{\text{Inc}} z \text{ and } Y \vDash_{\text{Inc}} z \tag{8.17}$$

- insufficient, if given Y, c does not grant inference to z

$$c, Y \nvDash_{\text{Inc}} z \tag{8.18}$$

The insufficiency case generalises the case of failure of external validity to inferences where z does not concern the obtaining of one relatum. The non-necessity case, instead, points to a problem of a different nature: given Y, c is a useless inference ticket. For instance, given the operation of gene 1, knowledge of the causal effect of gene 2 contributes nothing, or very little, to the inference to the insurgence of the disease.

8.1.5 Improving the causal picture

The procedure by which the community question and modify their causal beliefs is, roughly put, the following. One typically starts with some set of commitments implicitly shared with the other participants in the language game, and observes some failure in the inferences that he takes to be constitutive of the meaning of 'causes', which triggers a deliberation process as to whether the endorsed inference patterns are correct. A failure may be, for instance, the observation of the conditions described by the premisses but not of those described by the conclusion. However, unsuccessful predictions or actions are not the only trigger. Reasons of 'internal economy' in one's inferential activity, such as the easiness with which the inferences are drawn, or the usefulness of having a larger or smaller variety and number of inferences that can be drawn from the rest of our commitments, are other, legitimate grounds for revising the meaning of 'causes' (cf. de Donato Rodríguez and Zamora Bonilla, 2009a, pp. 106-107). In any case, the following choices are open to the participants in the game.

First, one may continue to endorse the inference, viz. take c as correctly applying to the circumstances in which the failure was observed to obtain. After all, a causal claim is not an inference license which gives a 100% guarantee of success, but a license that describes a modality which, although stronger than mere possibility, is weaker than necessity.

Secondly, one may blame the other premisses, and proceed to some substitution, addition, or shrinkage to restore correctness. This results in a more fine-grained specification of the conditions for the correct applicability of c. If the set of circumstances in which c is correctly assertible is shrinked, this may eventually and gradually lead to abandoning the commitment in the relation being causal.

Thirdly, one may blame c itself for the failure. This choice is motivated by the observation that any modification to the collateral premisses which restores correctness also results in c being unnecessary to the inference. One simple case is when the conclusion would follow if only c were removed, and the other premisses were left untouched.

In general, if commitment to the existence of the causal relation is abandoned, this is usually done on the ground that, given available knowledge, success and failure of the inferences which were taken as belonging to the meaning of 'causes' are best accounted for by modifying our causal picture, in any case by removing c from the arguments in which it used to figure. This should result in a global improvement of *all* the inferences.

As it happens with other claims and concepts, such as laws or kinds of superseded theories, the modification can be either local or global. Sometimes some local fixing is enough: one need not make major adjustments, only replace the claim in the arguments where it appears with one or more other claims, or simply remove it. At other times, the modification must be more global: claims and concepts to which the causal claim was inferentially connected are too dependent on each other, and must be substituted en masse. For instance, one may abandon the concepts that stand for the relata, if the relata's identity is too closely dependent on their role in the causal relation in question, and not enough on their role in other relations. In the latter case the abandonment of the causal claim may also drag the abandonment of the concepts that identify its relata.

Biological and economic kinds make interesting cases. Biological kinds are both functionally and structurally defined. For instance, to be a Bcl protein is, largely, to play such-and-such a promoting role with regard to apoptosis. So, finding out that the Bcl family contains not just proapoptotic but also antiapoptotic proteins may lead to abandoning the kind. However, 'Bcl protein' is also structurally defined. Genes that code for proteins in the Bcl family share a common coding sequence. The kind 'Bcl protein' is eventually retained, in spite of the failures of 'Bcl promotes apoptosis' to be

a good inference license, because the pay off of having the concept is larger than the advantage of getting rid of it. So, although the removal of a causal claim in biology can in principle determine a big shift in the meaning of biological kinds related by the claim, usually there are independent reasons why the kinds are entrenched enough, irrespective of the goodness of the claim. This is what gives us the impression that causal relations have a sort of second-order reality with respect to that of their relata. But due to semantic holism, the difference is only of degree. In principle, it is as possible to abandon the commitment to the existence of a relatum as it is to abandon the commitment to the existence of the relation. In both cases, the objectivity depends on the robustness of the meaning-constitutive arguments. So, if the sharing of a common coding structure were not sufficient, or not necessary, for the belonging to the Bcl class, or if the belonging to the Bcl class failed to be relevant, whether positively or negatively, to apoptosis, then probably the kind itself would be abandoned together with the causal claim.

This becomes even more evident in the case of economic kinds. These are often only functionally defined. As I argue below, model variables such as learning speed, being fundamentalist or chartist, network connectivity, etc. are harder to interpret in relation to the system, and may refer to quantities that have a precise meaning only in the model. This weakens the objectivity of the causal relations in which they take part. If the assertibility of the causal claims outside the domain of the model turns out to be too fragile, this makes the kinds related by them fragile, too. And if the claim is abandoned, the kinds are easily abandoned, too.

It is important to stress that which choice is eventually made is not something which is forced on the participants of the game. Since neither high-level rules of scientific reasoning nor low-level rules governing the meaning of claims and concepts obey a context-independent logic, but depend on material conditions, they cannot be interpreted as always leading to clear yes-no answers, so argumentation and deliberation never become dispensable. In Zamora Bonilla's words,

> individual behaviour is often not fully determined by commitments. There are many reasons for this indeterminacy: agents can be entitled to choose between several options; commitment usually comes in degrees; someone can be committed to perform two incompatible actions; and agents can decide to break their commitments sometimes. In all these cases, there is some room

for 'strategic' choice (Zamora Bonilla, 2006, p. 190).

I now turn to illustrate how to assess the objectivity of causal relations in systems biology and computational economics with reference to examples of causal claims from the case studies.

8.2 Causality in systems biology

What is special about the meaning of causal claims in systems biology with respect to the meaning of 'causes' *simpliciter*? To answer this question, one should investigate how the rules that govern the correct assertibility of causal claims are applied in systems biology. Systems biology is characterised by the joint use of difference-making and mechanistic criteria to establish causal claims, and the substantive weight of new methods, e.g. simulation, both to establish and to draw conclusions from causal claims.

In §7.2.3, I illustrated the meaning of '$C_{<\text{XIAP feedback, irreversibility}>}$' as established by Legewie et al. (2006), in terms of the inferential potential of 'XIAP feedback causes irreversibility of Casp3 activation', in short c_1. Below, I make explicit the meaning of '$C_{<\text{TNF, irreversibility}>}$' as established by Mai and Liu (2009). The authors investigate whether TNF causes irreversibility of DNA damage, that is, whether and to what extent the claim 'TNF causes irreversibility of DNA damage', in short c_2, is correctly assertible. They take evidence of difference making as warranting the causal claim, and the claim, in turn, as entitling to a number of qualitative, robust conclusions:

$\boxed{\mathsf{c}^{\downarrow}}$ TNF *makes a difference to* irreversibility and there is a plausible mechanism $\models_{\text{Inc}}$ TNF *causes* irreversibility

$\boxed{\mathsf{c}^{\uparrow}}$ *Predictions*: TNF_{ON} & apoptosis $\models_{\text{Inc}}$ irreversibility; TNF_{ON} & GF_{OFF} & survival $\models_{\text{Inc}}$ apoptosis. *Explanations*: TNF_{ON} & reversibility $\models_{\text{Inc}}$ I_{ON}; TNF_{ON} & stable survival $\models_{\text{Inc}}$ GF_{ON}

"I_{ON}" indicates XIAP feedback, as studied in (Legewie et al., 2006). More precisely, it stands for the correct functioning of the Casp3 inhibition of XIAP, which makes is possible for XIAP to establish an implicit positive feeback mechanism, responsible for the irreversibility of apoptosis.

8.2.1 Warranting the claim

In the absence of large databases to investigate regularities and/or probabilities, in the absence of clear intuitions about counterfactual scenarios, and in the absence of the possibility of surgical interventions, computation and simulation are more and more used to support causal conclusions. What is special about systems biology vis à vis more traditional molecular biology is that difference making established by means of computation and simulation constitutes evidence in its own right, not merely a useful heuristic device. Although systems biologists usually stress that *in vitro* and *in vivo* studies are necessary to establish their hypotheses, they also interpret *in silico* experiments as having a crucial role in *testing* such hypotheses (Kitano, 2002a,b; Westerhoff and Kell, 2007). This is acknowledged by philosophers of biology:

> Once a model has been developed to an appropriate level of complexity, it can be run repeatedly by a computer and function as a high-throughput hypothesis tester. Ultimately, the results of simulations must confront more traditional real-world experimentation, although the proportion of such tests reduces bench experimentation to a supplement or safeguard (O'Malley and Dupré, 2005, p. 1272).

On the one hand, models providing some *plausible mechanism* (or structure, or network, or the like) are usually necessary to back causal conclusions in systems biology. This is especially true of conclusions drawn with the aid of 'bottom-up' models, which make use of a greater level of mechanistic detail (cf. Bruggeman and Westerhoff, 2006), and c_1 and c_2 are no exception. But this observation must be carefully interpreted. Systems biologists often emphasise that the quantitative dimension is essential to the emergence of systemic properties. So, although the plausibility of the mechanism is necessary to establish a causal claim, it is far from sufficient. Knowledge of the amount and concentration of reactants, the affinities, the speed of reactions, etc. that make a difference to the obtaining of the effect is essential.

On the other hand, studies that establish their conclusions by producing difference-making evidence lend themselves more naturally to a causal interpretation. But here, too, some care is needed.

First, causal conclusions never come without some prior, causal assumptions. Without including a given causal relation in the base, or without as-

suming the existence of a plausible mechanism, the conclusion on the target causal relation would not follow. For instance, Mai and Liu (2009) model the regulatory node I, and study the dependence of apoptosis and survival on the state of this node, based on the assumption that Legewie et al. (2006)'s study resulted in the identification of XIAP feedback's causal role in establishing irreversibility. Mai and Liu (2009) find that setting I=OFF results in "noticeable degradation in the irreversibility of the apoptosis process" (*ibid.*, p. 765), but does not stop apoptosis. Without this and other causal assumptions, they would not be able to establish that TNF causes irreversibility.

Secondly, causal claims in systems biology are often established in the absence of the satisfaction of each and every condition postulated by traditional analyses. True, the evidence that supports the claim can consist of, or be interpreted as, a regularity, a probability-raising relation, a counterfactual relation, etc. Sometimes the evidence can be interpreted in agreement with all difference-making intuitions. At other times, only some intuitions are satisfied, not others. In any case, for difference-making evidence to grant the causal claim, the satisfaction of all conditions assumed by reductive accounts of causality is not necessary. For example: given the presence of Casp3 feedback, presence of XIAP feedback may be regularly associated with Casp3 irreversibility, and yet the association may break down upon expansions of the factor frame or changes in the underlying mechanism; presence of Casp3 irreversibility may counterfactually depend on XIAP feedback in one context, defined by the model where c was established, and not in another context, defined by another model where, e.g., TNF is ON. This is why *ceteris paribus* clauses are added, to alert that test criteria must be used *cum grano salis*.

In sum, the widespread use of difference-making evidence does not support the claim that what scientists mean by 'causes' is reducible to some difference-making criterion or other.

8.2.2 Using the claim

The criteria to establish the objectivity of causal relations are not limited to the criteria to assess whether the causal claim is warranted. Systems biology shows that assessing whether causal claims entitle one to certain conclusions is as important as assessing whether the causal claim itself is supported by the evidence or not. One may agree that a commitment to the causal claim would follow from a commitment to certain evidence, and take at face value

most of the causal conclusions found in the literature, but still question the endorsement of such conclusions on the ground of what does or does not follow from them. Pointing to cases where c does not warrant the inference to some z—which comprises the cases falling under the second and third categories mentioned in §8.1.4—is another way to challenge the objectivity of the relation. Correspondingly, ensuring that c does warrant the inference to a range of consequences is a way to vindicate the objectivity of the relation. The peculiarity of systems biology vis à vis traditional molecular biology is that novel tools are employed to this end.

Before turning to illustrate how these tools are used, it is instructive to compare the case of systems biology with that of economics. Cartwright (2007b, chap. 16) laments that the counterfactuals that are useful to economic policy are not the counterfactuals with which economists are typically concerned. Economists focus on establishing whether the cause makes a counterfactual difference to the effect in 'implementation-neutral' conditions, such as RCTs, or 'epistemically convenient' conditions, such as controlled experiments. Such counterfactuals, Cartwright argues, are informative on how much the effect variable would change in the case of implementation-neutral changes in the cause variable or in the case of modular systems, where the effect can be independently manipulated by changing the cause and keeping the background fixed. However, they are not so informative for policy, where one wants to know how much a cause contributes to the effect in non-implementation-neutral or non-epistemically-convenient conditions. Cartwright concludes that traditional methods in economics do not say much on this, in particular on "how our methods for hunting causes can combine with other kinds of knowledge to warrant the uses to which we want to put our causal claims" (*ibid.*, p. 175). She herself can only offer some general recommendation to the point that "the exact details matter" (*ibid.*, p. 241) and that one must ensure that the causal contribution carries over from test to target context (*ibid.*, p. 257).

Systems biology, in contrast to traditional economics or molecular biology, makes use of simulation to address these issues. In systems biology, the role of simulation to determine the consequences of a causal claim is even more relevant than its role for testing the causal claim. In fact, whereas testing requires evidence of the cause's capacity to make a difference in some implementation-neutral or epistemically convenient circumstance, using a claim requires evidence that the cause will make the right amount of difference in the target

circumstance. The latter kind of evidence is harder to gather than the former, whence the utility of simulation.

This, too, can be illustrated by reference to c_1 and c_2. These are claims that the systems biology community regards as well-established and that, arguably, do not suffer from problems of internal validity. Still, that may fail to robustly entitle one to certain consequences (see below) depending on what collateral premisses are endorsed. Whether or not they do entitle and what they entitle to is evaluated either by *simulating* the cell's possible behaviour in different contexts or, in the absence of a model to perform a simulation, by *imagining* (with the aid of background knowledge) what the most likely behaviour would be, which is a sort of simulation, viz. a thought experiment.

In particular, the robustness of $C_{<\text{XIAP feedback, irreversibility}>}$ depends on assumptions as regards $C_{<\text{TNF, irreversibility}>}$, and vice versa. As I explain below, conclusions with regard to certain consequences of c_1 would not be warranted were it not for the collateral commitment to c_2 and to the state of TNF being ON or OFF; nor would conclusions with regard to certain consequences of c_2 be warranted were it not for the collateral commitment to c_1, and the state of XIAP feedback, which is represented as either I_{ON} or I_{OFF}. Of the two conclusions, only the latter is actually arrived at by simulation; the former, instead, is drawn with the aid of knowledge on the causal role of TNF.

Let me begin with a failure of c_1 to successfully entitle to the presence of XIAP feedback, based on knowledge of irreversibility of caspase activation.

Example 1: $C_{<\text{XIAP feedback, irreversibility}>}$

1. Hypothesise: irreversibility & $c_1 \vDash_{\text{Inc}}$ XIAP feedback
2. Observe: irreversibility & no XIAP feedback
3. Diagnose: c_1 is insufficient; TNF_{OFF} is necessary, too
4. Modify: c_1 & irreversibility & $\text{TNF}_{\text{OFF}} \vDash_{\text{Inc}}$ XIAP feedback
5. Verdict: no entitlement to TNF_{OFF}

In the light of the observation, three choices are available, namely judge the argument OK as it is, blame c_1, and blame other premisses. In the circumstances, it seems reasonable to opt for the third choice: something more is needed for the entitlement to the conclusion. In general, entitlement to either bistability or TNF_{OFF} is necessary for the conclusion. If bistability

were present, then XIAP would arguably be observed. So, assuming the absence of bistability, the failure of the inference is best explained by reference to TNF. One can then modify the inference pattern accordingly, and blame the lack of entitlement to TNF_{OFF} for the failure. Since the inference turns out sensitive to premisses other than c_1, and the conclusion is less robustly inferrable based on c_1 and irrespective of other premisses, this decreases the objectivity of $C_{<\text{XIAP feedback, irreversibility}>}$.

A similar reasoning may be shown to apply to the case of a failure of c_2 to successfully entitle to I_{ON} based on knowledge of TNF_{ON} and reversibility.

Example 2: $C_{<\text{TNF, irreversibility}>}$

1. Hypothesise: c_2 & TNF_{ON} & reversibility $\models_{\text{Inc}} \text{I}_{\text{ON}}$
2. Observe: TNF_{ON} & reversibility & no I_{ON}
3. Diagnose: c_2 is insufficient; Casp3 feedback is necessary, too
4. Modify: c_2 & TNF_{ON} & reversibility & Casp3 feedback $\models_{\text{Inc}} \text{I}_{\text{ON}}$
5. Verdict: no entitlement to Casp3 feedback

Here, too, one may decide that the argument is good as it is, blame c_2, or blame other premisses. In the circumstances, one may reason as follows: knowledge that Casp3 feedback is present was necessary, too, for the conclusion. However, one was not entitled to this additional premiss. Also in this case, the fact that a conclusion turns out less robustly inferrable from the causal claim makes the relation less objective.

The above examples show that the semantics of different relations may depend on the interplay between the corresponding claims, since one claim can appear in the base, or the collateral commitments, or the target of the other. So, a claim which is established in one context can serve as a reason for another claim in another context.[122] This is in line with the inferentialist idea that, due to semantic holism, endorsing or rejecting *one* commitment is often a matter of endorsing or rejecting *other* commitments.

In sum, in systems biology, too, causal claims can, and do, serve the function of inference licenses. However, whether the license leads to successful

[122] In this regard, notice that there is no contradiction in claiming that the existence of one relation reduces the local objectivity of another and that the two relations together increase objectivity globally. It may be that, although the objectivity of *one* relation decreases if the assertibility of the corresponding claim is sensitive to another relation, postulating the *two* relations allows that the two claims mutually reinforce each other, so that the set of inferences in which the two appear is *globally* strengthened.

uses depends on—possibly many—other commitments. In general, the correct assertibility of causal claims in systems biology tends to be very sensitive to the context of their application. This is due not only to the complex character of the context itself, but also to the ability of the model to work as a good surrogative tool for reasoning about that context. As a result, the sensitivity by which causal claims contribute to the correctness of the arguments where they figure makes the relations referred to in the claims less objective, too.

8.3 Causality in computational economics

Let me now come to the meaning of causal claims in computational economics, which I illustrate by reference to the meaning of '$C_{<\text{switching, volatility}>}$', made explicit in terms of the inferential potential of 'switching causes volatility' (c_3) (Lux and Marchesi, 1999, 2000), and the meaning of '$C_{<\text{learning speed, volatility}>}$', made explicit in terms of the inferential potential of 'learning speed causes volatility' (c_4) (Arthur et al., 1997; LeBaron et al., 1999).

8.3.1 Warranting the claim

I will start by discussing the downstream potential of causal claims in computational economics. The downstream potential of c_3 and c_4 is, respectively:

$\boxed{\mathsf{c}^{\downarrow}}$ switching *makes a difference to* volatility $\models_{\text{Inc}}$ switching *causes* volatility

$\boxed{\mathsf{c}^{\downarrow}}$ learning speed *makes a difference to* volatility $\models_{\text{Inc}}$ learning speed *causes* volatility

Causal claims involving causes of crashes in the stock market seem less warranted than the claims involving the causes of apoptosis. Why? This has to do with the conditions for being committed to there being warrant for the claims.

Both models answer the question 'what causes volatility in the stock market?', and more generally, 'what causes the stylised facts?' in terms of the *heterogeneity* of the agents in the market and their *interactions*. These two factors give rise to an endogenous self-reinforcing process, viz. a positive feedback not adequately counterbalanced by a negative feedback. One model describes the mechanism as based on a self-reinforcing fundamentalist-chartist switch, driven by imitation. The other model describes the mechanism as

based on chartist strategies becoming mutually reinforcing, due to the observation of prices and the others' profits. In either case, the outcomes are high volatility, (temporary) bubbles and crashes, and other stylised facts. Causal responsibility for these phenomena is ascribed to, respectively, *irrationality* (i.e. non-fundamentalist behaviour) of a high proportion of agents at some time, or *reflexivity* (i.e. inductive-adaptive behaviour) of all agents at all times. While these analyses are instructive as to what commitments would generate commitment to c_3 or c_4, it is not clear in what conditions one would be *entitled* to assert c_3 or c_4, since it is not clear how one can become entitled to the *warrant* for c_3 and c_4.

Consider $C_{<\text{switching, volatility}>}$ first. The problem here is not so much with the assumption that the stock market is chaotic. One may reconstruct the shape of a chaotic attractor, thereby providing some evidence that the stock market is chaotic, even in the absence of tests on a precise theoretical model[123]—which is hard to produce when we don't know what the relevant observables are. And one can take scaling laws as providing further evidence for the underlying chaos. Nor is the problem with the assumption that quantities be continuous, necessary for the sensitivity to the initial conditions. One may admit that this aspect of the model misrepresents the system, whilst maintaining that it correctly captures the 'stretching and folding' of the dynamics in a confined region, which is essential to chaotic behaviour (Smith, 1998, chap. 3).

Rather, the problem is with the mechanism which is supposed to generate chaotic behaviour and scaling laws. The mere isolation of scaling laws in the data may constrain the inference to the underlying mechanism (e.g., by excluding processes that generate Gaussian distributions) but is not sufficient for it (Rickles, 2011, §9).[124] What we need is evidence that a plausible mechanism has been identified. In the case of Lux and Marchesi's model, however, it is not clear what the commitment to there being difference making between switching and volatility amounts to. Although at some abstract level it makes sense to talk of more or less chartist or fundamentalist behaviour, at a more concrete level it is harder to figure out what the switching from being chartist to being fundamentalist is like. Ultimately, the warrant for c_3 depends on the

[123]To this end, one can take the time series of one observable, displace the time value to obtain two or more 'fake' observables, then plot them in space and study their evolution to see whether they approximate the shape of a chaotic attractor (Stewart, 1997, pp. 172-178).

[124]Nor is the theory of self-organised criticality sufficient to demonstrate the existence of such a mechanism if, as some maintain (cf. Frigg, 2003), this is only a group of models united by a 'formal analogy'—which, as such, don't explain.

goodness of the analogy between switching in the market and *phase transition* in many-particle systems, and this analogy has both virtues and limitations.

A virtue of the analogy is that in both cases the individuals' *interaction* is crucial for the emergence of the behaviour. Lux and Marchesi's model captures an important aspect of human decision making: traders change their strategy based not only on the observation of price series but also on imitation, here modelled as observation of moods and profits of the other traders. However, the analogy has also limitations. Contrary to particles, agents are 'intelligent'. The model does not open the black box of the process of formation and evolution of their forecasting strategy. More importantly, it treats agents as neatly groupable as particles in different states, whereas the representation of their heterogeneity may need more fine-graining. The model postulates the existence of unintelligent behaviour and fully rational behaviour, which is obviously a gross simplification and idealisation. On the one hand, also 'unintelligent' noise traders should learn from 'mistakes'. However, learning should be conceived not as a switch to a fully rational behaviour, only as an attempt to improve on actual strategies not to incur in the same forecasting failures. On the other hand, one's intelligent trading should not be identified with full knowledge of the expectations of both noise traders and fundamentalist traders, one's own expectations included. This is because if heterogeneity amongst traders is assumed, the use of a deductive expectation formation process would generate a regress (cf. §3.2.2).

As far as the main goal of the model is concerned, these details may not matter much. The model only aims at showing that non-classical conditions, if instantiated, would be sufficient for the emergence of the stylised facts. Yet, in the absence of a strong analogy between the mechanism underlying bubbles and crashes and the mechanism underlying phase transition, it is hard to identify the cause of the macrobehaviour, which is a necessary condition for the transfer of warrant from premisses to conclusion to take place. A symptom of this is that, contrary to the many-particle case, in the stock market the quantities postulated by the model have no clear counterpart. As a result, it is hard to become entitled to c_3.

In the case of $C_{<\text{learning speed, volatility}>}$, too, it is hard to become entitled to the assumptions in the model. The Santa Fe model explains stylised facts not in terms of phase transitions, but in terms of *evolutionary changes*. Under the assumption of the agents' heterogeneity and bounded rationality, changes in market regime are explained as the emergence of mutually-reinforcing induc-

tive behaviours ultimately produced by learning and adaptation. However, the Santa Fe model, too, has problems.

Some of these problems are more of a technical nature, e.g., the assumption that wealth does not affect share demand, the lack of a quantitative match of the results to actual financial data, etc. Another problem is the use of the classifier system to map past information into trading strategies. Although this is a useful metaphor for learning, it is clearly not so realistic.

Even on the assumption that the above problems may be fixed, or are not significant, there are more significant limitations. First, the model does not account for the fact that learning agents are capable of finding an equilibrium, by adjusting their learning pace. The model cannot explain this *coordination* process. In the model, there is no social learning between agents, the only learning is via the observation of prices. Secondly, changes in regime depend on a unique parameter, viz. learning speed, whose role in terms of real market mechanisms is unknown. LeBaron himself, one of the creators of the Santa Fe model, observes: "If a single parameter for which we know little about in reality can change the outcome so dramatically then we may always be in a state of uncertainty concerning potential model predictions" (LeBaron, 2002, p. 14). Giving a realistic account of the agents' process of expectation formation is—notoriously—extremely complicated (cf. Simon, 1996, pp. 36, 39). This is not to say that the Santa Fe model says nothing interesting. On the contrary, it does identify in the agents' process of expectation formation a crucial determinant of volatility. However, without a better grip on how this process is instantiated, it is hard to become entitled to a claim whose assertibility depends on such a process.

In sum, the phase transition analogy and the evolutionary biology analogy give at best a partial story. Neither Lux and Marchesi's model nor the Santa Fe model provide a particularly plausible psychological mechanism. Besides, we know that the heterogeneity on which their results depend may concretise in many different ways, not just as a difference in forecasting strategy with regard to the next period's price and dividend, but also as, e.g., a disagreement on the time required for the price to converge to the fundamental value, an asymmetry of knowledge about the fundamental value, a difference in the investment horizon for different investors, etc. (Markose et al., 2007). Finally, neither model accounts for the obvious causal relevance of the network structure proper of the stock market, which is arguably characterised by agents of *different kinds*—not only individuals, but also firms, banks, regulatory insti-

tutions, etc.—and *non-symmetric interactions* (Thurner et al., 2010). So, we have both positive and negative reasons to believe that we may not become entitled to the premisses that would entitle us to the correct assertibility of c_3 or c_4.

More instructive is to take the two models in combination, as jointly supporting the more general conclusion 'systemic instability causes volatility', where "systemic instability" stands for any endogenous mechanism responsible for destabilising the system, e.g., switching, inductive learning, some feature of the network's 'connectivity' (Kirman, 2010), etc. Although the two models describe very idealised mechanisms, they show that also in the absence of external shocks (big changes in fundamental value) volatility and crashes obtain robustly, not only upon relaxation of assumptions in the model (variations in parameter values within the same mechanism)[125] but also upon variation in the underlying mechanism (distinct mechanisms lead to the same results)[126]. In this case, it is the two models themselves that entitle us to the conclusion that systemic instability—no matter how instantiated—causes high volatility and crashes:

$\boxed{\mathsf{c}^{\downarrow}}$ switching *makes a difference to* volatility; learning speed *makes a difference to* volatility $\models_{\text{Inc}}$ systemic instability *causes* volatility

This way of using the two models is in line with the first criterion for objectivity in 8.1.1, viz. independent derivations drawn from distinct sets of assumptions contribute to the robustness of a conclusion. If distinct models, relying on different assumptions as regards the nature of the heterogeneity of the individuals, support the same conclusion, this is good evidence that the conclusion is plausible.[127]

8.3.2 Using the claim

A separate issue is whether $C_{<\text{switching, volatility}>}$ and $C_{<\text{learning speed, volatility}>}$ are objective in the sense that the *conclusions* drawn from c_3 and c_4 are robust. Among the inferences granted by c_3 and c_4 are, respectively:

[125] Behind this procedure is the maxim that Smith labels 'trust the robust', that is "take as seriously representational what is reasonably stable as precisification vary" (Smith, 1998, p. 128).

[126] This interpretation of model building is in line with Wimsatt's idea that even unrealistic, or 'false', models can be means to 'truer' theories (see Wimsatt, 2007, pp. 100-106).

[127] Indeed, there is evidence that similar outcomes are produced by a variety of models which make more realistic assumptions either on the nature of the heterogeneity of the individuals (Follmer et al., 2005) or on the market's network structure (Anand et al., 2011).

$\boxed{c^{\uparrow}}$ *Predictions*: sensitivities > threshold $\models_{\text{Inc}}$ volatility; sensitivities < threshold & chartists > fundamentalists $\models_{\text{Inc}}$ volatility. *Explanations*: volatility & random change in FV & sensitivities < threshold $\models_{\text{Inc}}$ switching; volatility & sensitivities < threshold & no switching $\models_{\text{Inc}}$ large change in FV

$\boxed{c^{\uparrow}}$ *Predictions*: fast learning $\models_{\text{Inc}}$ volatility. *Explanations*: volatility & random change in FV $\models_{\text{Inc}}$ fast learning; volatility & slow learning $\models_{\text{Inc}}$ large change in FV

Here the contribution of c_3 and c_4 is evaluated in terms of their role as premisses of arguments. The problem with drawing consequences—especially predictions—from c_3 and c_4 is that one is entitled to such consequences only provided he is entitled to the premisses, and for this to be possible the set of premisses must be 'appropriate', i.e., such that there *are* circumstances in which one can become so entitled. Two issues arise. First, it is hard to imagine how one could become entitled to the premisses involving the parameters that measure the putative causes, viz. switching and learning speed, taking on some value. Secondly, if one does not rely on the possibility to become entitled to such premisses, it is hard to find some suitable, additional premiss to be added to c_3 or c_4 to derive z from the other collateral premisses.

The first issue concerns the possibility to interpret the quantities in the model with respect to the target. Since it is hard to concretise the idealised assumptions in the model, it is also hard to use the model to draw conclusions with regard to some target (cf. de Donato Rodríguez and Zamora Bonilla, 2009a, p. 114). Yet, there are *other* inferences whose correctness does not rely on the entitlement to the particular state of the switching process or the agents' learning speed, only on the *existence* of the corresponding mechanisms. Here, although one may have no particularly good reason to describe a particular situation by means of c_3 rather than c_4, there are consequences that follow robustly from c_3 and c_4 *irrespective* of the exact details of the mechanisms behind $C_{<\text{switching, volatility}>}$ and $C_{<\text{learning speed, volatility}>}$. Arguably, in fact, if c_3 and c_4 can be derived from the models, this is only because exogenous shocks aren't strictly necessary to produce volatility, an endogenous process being in principle sufficient. This is enough to at least use c_3 and c_4 to cast doubt on claims such as 'Crashes are necessarily caused by exogenous shocks'. Let us indicate with "c_i" any claim to the point that systemic instability is sufficient to generate high volatility, e.g., c_3 or c_4. Then, at least one backward-looking inference is possible to whose premisses we may become

entitled, so that the we are also entitled to the conclusion:

$\boxed{\mathsf{c}^{\uparrow}}$ *Explanation*: c_i & crash & random change in FV $\models_{\text{Inc}}$ systemic instability.

The second issue concerns the possibility to add to the premisses a claim on how to measure instability without relying on the possibility to concretise the specific mechanisms postulated by the models. One way to do this is by means of *another* mechanism, for instance one based on an interpretation of the network's connectivity in terms of some measurable feature of the target system (Anand et al., 2011), so as to open the black box of the mechanism generating instability in the target system. Another way is by measuring some 'surface' feature of the—partially unknown—mechanism for volatility, e.g., the acceleration of price variation near the critical point as measured by the time-to-failure analysis (§1.3.4), so as to use the existence of the mechanism as a black box. If one or the other strategy succeeds, then also some forward-looking inference may become possible of the form:

$\boxed{\mathsf{c}^{\uparrow}}$ *Prediction*: c_i & random changes in FV & systemic instability $\models_{\text{Inc}}$ crash.

Obviously, it is very hard to become entitled to such an inference, but we should not think it is impossible. What one needs are reasonable—although defeasible—reasons. Such reasons may come in the form of both *positive* reasons (viz. collateral premisses) to which we may be entitled, and are entitled in the circumstances, and *negative* reasons, to which we may also be entitled, but are not entitled in the circumstances. The resulting complex of positive reasons and lack of negative reasons should make it more plausible to infer 'crash' rather than 'no crash'. Notice that the correctness of the inference depends not on the 'sameness' between the system in which the claim was tested and the system to which it is exported, or on the 'resemblance' between model and target system, but on the external validity of the claim.

Although the objectivity of the relations is not construed in representationalist terms, whether a relation is objective or not, and to what degree, is non-arbitrary. An analogy between, say, a crash and a phase transition phenomenon may partly be the result of a 'creative' act. We do know in advance that a good theory of crashes need not have the same inferential role as a good theory of phase transition, in the same way that a good model of the Phillips curve, such as a fluid mechanics model, does not have the same inferential role of the phenomenon it aims to represent. There are, in

fact, inferences that are correct when applied to one, but turn incorrect when applied to the other. However, there is also an objective component to the analogy: whether the success-conducive use of the notion of 'instability' to predict transitions in many-particles systems can be extended to the stock market case is largely an empirical matter, which depends on the inferential fruitfulness of the analogy for the causal claims in question.

However, as shown, the analogies at work in the above examples have significant limitations. As a result, with respect to the causal claims in systems biology, causal claims in computational economics (at least the ones considered here) tend to score lower in terms of assertibility: it is harder to become entitled to the premisses that warrant their application or to the consequences of their application, or the circumstances in which we may become so entitled are fewer. Since the range of circumstances in which the claims contribute to correct inferences is smaller, the objectivity of the causal relations described in such claims is weaker.

8.4 Grounding objectivity in normativity

For the representationalist-inclined, whether a causal claim is objective depends on whether it represents the world as it is, or its truth conditions are satisfied. For the inferentialist, instead, reference to the 'world as it is' (outside the space of reasons) and to truth-makers as unexplained explainers plays no substantial, explanatory or justificatory role for the correctness of our assertions. Whether a causal claim is correctly assertible depends on the grammar, or meaning, of the word 'causes' as well as the other words in the claim (the claim's or the model's assumptions are ones we are strongly committed to, for reasons of, e.g., internal coherence), together with some privileged epistemic role of observations and actions (e.g., predictions corresponds closely to our observations) (see de Donato Rodríguez and Zamora Bonilla, 2009a, p. 109). The objective status of a claim depends on the harmony which results from the (normative) process of challenge and revision of commitments. Applied to science,

> inferentialism does not attempt to find out what [the] 'right' ways of doing science are, but it helps to justify the objectivity of science by making us recognise that the epistemic claims to which we are finally committed are not necessarily the ones we would have wanted to defend in the first place, but the ones our reasoned

> dialogues have led us to accept in the end (Zamora Bonilla, 2006, p. 198).

When deciding on endorsing a certain causal claim, we employ entrenched rules (SIM), that is, the most robust rules that have survived the process of challenge and revision, and treat case by case problematic claims, to which such rules do not apply so well. Claims to which several, or all, test criteria apply, and from which a broad range of conclusions follow, will be more straightforwardly accepted. Claims such as those involving complex systems, which satisfy test criteria less strictly, and are conducive to less successful conclusions, trigger a deliberation process that involves both epistemic and semantic considerations. That is, we do not only ask how successfully the claim is confirmed (how many criteria are satisfied and to what extent) and used (how many correct predictions, interventions, explanations the claim makes possible). We also ask how our concept of causality is and should be used (whether and why we should apply it to the problematic context, too). Indeed (for the semantic holist) the two kinds of consideration cannot be neatly distinguished.

Is this enough to guarantee the objectivity of the causal relations referred to in the claims, irrespective of what is *actually* inferred by the speakers? The above seems at most a sort of 'coherentist' criterion, and one may still question the ability of the account to distinguish between an assertion that *seems* correct and one that *is* correct. Isn't there something more to objectivity than intersubjective agreement? It seems, so the objection goes, that the account needs—and does not have—a sort of 'independent handle' on correctness, that is, a way of *adjudicating* what assertions are correct—and not merely pointing to the existence of a meaningful distinction between what is correct and what seems to be correct. It is evident that relying on inferences *actually drawn* by an entire community, on the basis of criteria arrived at by conventional agreement and evolution, won't do. In fact, the satisfaction of these criteria may be insufficient (the criteria may be only accidentally satisfied) or unnecessary (the criteria's authority may be later questioned, on the ground that they are not exhaustive, or not consistent, etc. (see Zamora Bonilla, 2006, pp. 194-195)). Alternatively, one may appeal to inferences that *ought to be drawn*, or inferences which are 'better' than the actual ones. But how can one specify what count as better/worse inferences without an independent handle on what is correct?

Needless to say, addressing this debate in a satisfactory way would require an argument which goes beyond the scope of this thesis. Here, I only want to point to some reasons why the inferentialist should avoid to answer the question of what grounds the objectivity of causal claims by appealing to independent handles and standards of correctness external to the practice. I will do this by clarifying how the inferentialist should understand the nature of the relation between normativity and assertibility, and the standards of correctness by which to evaluate (systematic) mistakes. The resulting notion of correctness, I maintain, is good enough to understand the objectivity that is aimed for in science when, e.g., establishing and using causal claims.

First, what is the relation betweem normativity and assertibility? On the one hand, the inferentialist cannot but admit that one ought to obey the *actual* rules of the game when one infers the causal claim and uses it to infer other claims. On the other hand, the inferentialist maintains that one may be justified in refusing to obey a rule on how to use 'causes' on a given occasion, if he provides a good reason. Assertibility depends on a complex interaction between language users and the world, in which linguistic rules play a regulatory role, fully fixed by neither the world nor us. For instance, Brandom wants to allow for a notion of correctness which is *attitude-transcendent* (see Brandom, 2000, pp. 189-190). To this end, he relies on his 'normative fine structure of rationality', that is, the space of reasons that commitments and entitlements institute. The idea is that, once instituted *by* the speakers, this structure acquires a sort of autonomy and authority on matters of correctness *on* the speakers who instituted it (see Brandom, 2000, p. 203). How should we understand this? It is instructive to refer to a notion recently introduced by Peregrin (2012), viz. the notion of 'true normatives', something in between indicative sentences—which provide and stand in need of reasons—and normative sentences—which establish commitments and entitlements. We may think of causal claims as a sort of true normatives.

Although *all* meanings are normative, some seem 'more normative' than others. Analytic claims are more on the descriptive side, expressing (widely accepted) rules on the use of logical vocabulary or of notions that have simply been defined as "such-and-such" by stipulation. Observational claims are close to analytic in the sense that, due to the uniformity of human physiology, the use of the words involved is relatively uncontroversial. The more one moves to the other end of the spectrum, the more one finds normatively-loaded claims, involving theoretical, modal and moral vocabulary. The class

of true normatives lies somewhere in the middle. For instance, thick moral claims such as 'This is cruel' may count as true normatives. On the one hand, 'This is cruel' describes facts about deep suffering as well as the speaker's disapproval, so that inferences such as 'If such-and-such is done to somebody, it is cruel', and 'If this is cruel, I won't do it to anybody' are meaning-constitutive. On the other hand, 'This is cruel' contributes to enforce inferences such as 'If this is cruel, you ought not do it to anybody'. Analogously, but to a lesser extent, also causal claims such as 'XIAP feedback causes irreversibility' or 'Systemic instability causes volatility' have a normative component: if one is committed to the claims, one ought not deny that some predictions, explanations and interventions follow from them. When predictions, explanations and interventions fail, one ought to find what additional commitments may be compatible with such a failure before giving up on using the causal claim. This is not to say that true normatives are conventional, and one is at liberty to endorse or reject them no matter what:

> First, normatives must be anchored in the existing practices with existing rules and though they may, and usually do, go beyond them, they can do so only to such an extent that it makes sense to say that they are exercisings of the existing rules. Second, if a normative aims at a modification or an extension of the existing practices, it counts as a proposal, which can be taken as established only if it survives any occurring criticism and if it comes to be generally accepted (Peregrin, 2012, p. 94).

So, one should obey existing rules for the use of "causes", e.g. SIM. If one wants to change the rules, one cannot modify them arbitrarily but must provide reasons why specific contexts demand that one neglect them. Such reasons will involve not just evidential considerations (on why certain test criteria are more relevant than others, etc.), but also changes in opinion regarding core inferences (on why some test/target criterion should be added, or eliminated, etc.) which are made necessary to maximase the efficiency and the coherence of our *other* commitments.

True normative are 'true' in the sense that,

> though we can say that this stepwise development of rules amounts to a creation, it is usually understood as a case of *discovery*. To explain this peculiar feature of rules, compare their status with

> the status of certain objects of mathematics. On the one hand, it is acceptable to say that it was Cantor who devised sets, or that it was Galois & comp. who invented groups. But on the other hand, these mathematical objects were devised as entities which exist timelessly, and as such *cannot* have been *brought* into being. Thus, once we devise them, we must look at them as having been *discovered* by us, while having been here all along (*ibid.*).

When certain rules are in place, one's choices regarding the applicability conditions of a concept and the possible modifications of the rules that govern its correct use are constrained. In this sense, norms—scientific norms included (cf. Zamora Bonilla, 2006, p. 197)—are both *instituted by us* and *binding us*, like "the principles formulated by judges at common law, intended both to codify prior practice (...) and to have regulative authority for subsequent practice" (Brandom, 2000, p. 76). This seems the strongest notion of objectivity open to the inferentialist. If one feels he needs another, stronger story, perhaps he must look elsewhere. Be that as it may, this notion of objectivity seems strong enough for the scientists' needs.

Let me now come to the second point: What are the standards of correctness for adjudicating systematic mistake? Where do they come from? Correctness, for the inferentialist, does not presuppose a perspective from outside the practice, as the advocate of truth-conditional semantics has it, or some global perspective on the scorekeeping practice (a group of experts, or the whole community) under actual or ideal conditions (see Brandom, 1994a, pp. 594-595, 600-601). Questions such as 'Is our whole conceptual apparatus appropriate?' or 'Is the causal claim *really* true?' are the expression of a view from nowhere. And since we are somewhere, not nowhere, the inferentialist does not (should not) attempt to answer these questions, on pain of giving up on his inferentialism altogether.[128] In fact, any answer to these questions would make the inferentialist position *descriptive* after all: there's a language- or practice-free standpoint (ideal conditions, dispositions, or what have you) from which to assess correctness.[129] The next question is then:

[128] This is, of course, not to say that the questions are meaningless. However, what their meaning is, whether they need answering, etc. is another story.

[129] Notice that Brandom (1994b) explicitly rejects the interpretation of correct inference as the inference one would be disposed to draw in ideal conditions. So, if his position is bad, this must be for some other reason, e.g., he's unable to make sense of correctness without appealing to ideal conditions. Brandom's story on how dispositions get into the picture is that they are crucial to grant certain claims, e.g. observational reports, but are only *prima*

Why bother with practice and not just talk about the thing itself? Outcome: inferentialism is unnecessary, let's go representationalist.

So, the standards of correctness which adjudicate systematic mistake must be construed as *internal* to the practice. But how exactly? Here is a fruitful way to address the issue. If we construe the standards of correctness as internal to the practice, the relevant question to ask is: *whose* practice? And the question must be answered from within *some* practice, by appealing to sets of rules that indicate how to challenge each other for our failures to latch on to the world. It is only from within a practice that the issue of objectivity acquires sense. In Zamora Bonilla's words,

> evaluating a scientific argument from an external perspective amounts to putting the following question: what would have been a more correct way of doing it, according to *you*? (Zamora Bonilla, 2006, p. 198)

Two are the cases: (1) a language user, or community, is systematically wrong in its use of the concept; (2) the whole community is systematically wrong in its use of the concept.

In case (1), we do have communication between two or more parties; the parties utter the same words (e.g. "causes"); and they disagree on the rules of application of the word (on whether a causal claim is well established, or can grant other claims). Two possible outcomes: either (i) the parties agree they are using different concepts after all, or (ii) they agree they are using the same concept but disagree on the rules of its application. (ii) is the most interesting case. Reaching agreement here amounts to renegotiating the meaning, by weighing reasons against each other's commitments. Mistake can only be evaluated ex post, from the point of view of the reached agreement: one party wrongly took itself as committed or entitled to the claim, but the material inferences on which it implicitly relied were wrong. (If no party admits the mistake, this is an instance of (i).) Commonly this is not an all-or-nothing affair: both parties end up revising their sets of commitments.

facie reliable: although they generate quite uniform responses, they are still subject to a default-and-challenge model of belief revision. Meaning of no concept, e.g. 'red', is reducible to assertibility in ideal conditions. This would presuppose some ideal standpoint from which to assess how exactly "red" contributes to the truth conditions of "red-" claims. Which in turn presupposes the possibility to spell out the *ceteris paribus* clauses that accompany inferences to and from "This is red" so as to make such inferences *necessarily correct*. Requiring this much, for Brandom, amounts to trivialising the notion of objectivity.

In case (2),there is no communication *between* parties—there are no disagreements on the use of the same concept, so there are no parties as such. Also here I see two possibilities: (i) nobody is aware of using the concept differently from the others, and although there is a lot of talking of "x", some massive coincidence makes it possible that there is no disagreement on 'x'; more realistically, (ii) the whole community think to be using the concept in the same way and correctly. Either way, the issue is: What are the standards of correctness to blame the whole community? That is, on what principled ground can one say everybody is wrong, whether for different reasons of for the same reason? The inferentialist can offer no language-free standpoint to answer this question. Actually, he will say that he does not understand what the question means: As soon as you ask it, you side with some against others, or believe you're right and everyone else is wrong (cf. the above quote from Zamora Bonilla). This presupposes we are, after all, in case (1).

Conclusion

In this chapter, the inferentialist account developed in chapter 7 was applied to the analysis of causal claims in complex systems sciences. I discussed how the inferentialist can make sense of the objectivity andthe referential function of such claims in terms of the normative conditions that make for their assertibility. I characterised the objectivity of causal claims in inferentialist terms. I then illustrated my proposal with reference to claims on the causes of, respectively, apoptosis and asset prices' volatility. The sense in which the identified causal relations are objective was examined along the two distinct dimensions of warrant for the claim and warrant for its use. Finally, I discussed in what sense the inferentialist can claim to have offered a notion of objectivity which is attitude-transcendent, and suitable to capture the notion of objectivity at work in scientific practice.

Conclusion

Complexity calls into doubt the adequacy as well as the relevance of our armchair intuitions about causality. More specifically, it makes it implausible to analyse causality in terms of fully objective, mind-independent facts, and demands that we investigate the meaning of causal claims in a way which is more faithful to the practices where they are produced and used.

A satisfying account of causality in complex systems should, for one thing, explain how causal talk is employed by the scientists themselves in their specific areas of inquiry, and for another, provide some story on how *their* notion of causality is related to *other* notions of causality, as employed in other areas of discourse.

Inferentialism allows one to do just that. It reverts the traditional order of analysis, by taking our activities of agents as the raw material in terms of which to account for the obtaining of causal relations. Causality becomes a 'category' that the knowing subject employs to 'mediate' between himself and the world. In inferentialist terms, this mediation is the result of the concept of cause figuring in a network of inferences, used in our practice of gathering evidence and using it to explain, predict and intervene.

The meaning of 'causes' as relative to specific relata is made explicit in terms of the inferential potential of the claims where the causal relation is described, that is, in terms of the claims' contribution to the correctness of the arguments where the claims figure as premisses or conclusion. By an analogous reasoning, the meaning of 'causes' *simpliciter* is made explicit (whether in complex systems or other areas of discourse) as an inference license which can be typically inferred from a given cluster of criteria and which in turn entitles to inferences involving the obtaining of the relata.

By means of this process of explication, inferentialism promotes the identification of the contexts which we implicitly take to warrant the inference to the claim, and the jobs that we implicitly want the claim to do for us, that is, the entitlements that we expect the claim to provide us with and the purposes that we expect the claim to help us satisfy. In this way, inferentialism can both account for variability of the meaning of 'causes' depending on context

and subject matter, and encourage discussion on the one set of inferential rules that all speakers ought to follow when using the word "causes". Competent speakers, in fact, can/should ascribe to each other commitments on the same subject (e.g., a causal relation), praise and blame their inferential success on the basis of such commitments, and refine their own set of commitments accordingly, by (among other things) removing inconsistencies that become manifest along the process.

As evidenced by the study of complex systems, contrary to 'ordinary' concepts whose meaning is more stable and easily identifiable, 'causality' is a peculiar concept. In complex systems, extra-linguistic success granted by the application of causal claims and intra-linguistic agreement on the rules of their correct application become harder to get. By challenging our ability to achieve such goals, hence by making the mediating role of 'causes' more difficult, complexity also helps highlight its status of 'contested' and context-sensitive concept. The context-sensitive nature of causality makes the inferentialist approach particularly appropriate—*more* than in the case of other concepts. Inferentialism, in fact, allows one to formulate a flexible analysis, without at the same time endangering its informativeness.

Should one be a *global* or only a *local* pragmatist? Relatedly, should one analyse all concepts inferentially, or only the concept of causality? In my view, one may consistently hold that pragmatism about meaning is an attitude that should be endorsed globally, whilst maintaining that an inferentialist semantics is best regarded as a *tool* and, as such, as more or less *useful*. Reductive analyses may be (locally) more or less successful. Whenever they prove less successful—as is the case of, e.g., causal, modal and moral vocabulary—one should better turn to inferentialism to provide a more adequate analysis.

One reasonable question to ask is: *On what grounds* can an inferential analysis of 'causes' be correct? Is the (dis-)agreement among scientists or philosophers enough to settle the issue? The answer is a qualified 'no'. Science provides a material that needs interpretation. By looking at science, first and foremost we acknowledge that science may have more-fine grained tools to understand reality so as to achieve goals that are common to the laymen, too. In this process, both scientists and philosophers should make explicit the *connections* between tools and goals. This results in a continuous clarification of what we mean and/or ought to mean.

What about the observation that we do have different intuitions, after all, and constantly produce counterexamples against one another's analyses?

What moral should we draw from this? Now, failures and counterexamples could be interpreted in several ways. Sometimes they show that the question is wrongly-headed, or that we should change 'research programme'. When 'causality' is concerned, the central role that the notion plays in our conceptual apparatus demands that we don't dismiss the question. As long as the purposes with which causal talk is associated are stably related to one another, then it makes sense to ask what it is that makes them so. At the same time, the troubles of available accounts to deal with complexity does motivate a sort of change of research programme. Inferentialism provides the resources to re-interpret an intuition or a supposedly exhaustive analysis into *one* of the premisses from which the causal claim can be 'typically' inferred and/or *one* of the conclusions which can be 'typically' inferred from it. So the issue of whether or not the analysis is correct translates into the issue of understanding the *weight* of one criterion against another in a particular case (e.g., the counterexample) allowing that *all* criteria may be 'typically' inferentially related to the concept of causality.

The picture that emerges is one where the concept of causality is more dynamic and flexible than philosophers use to think. It takes on different nuances in different domains, and adapts to the features of the phenomena by adjusting the weight of the criteria that constitute its meaning. The issue of what 'causes' means cannot be settled once and for all, by either scientists or philosophers. If our concepts are dynamic, only partly constrained by our practices and Nature's inputs and outputs to such practices, we can only try to interpret concepts on-the-fly. Their meaning—'causes' included—must be called into question, made explicit and renegotiated, in a never-ending 'virtuous' circle. There is no ultimate court and no ultimate judge. In the case of causality, one (I) can only hope that meaning can be made explicit *enough*, core inferences identified, and vagueness contained.

Bibliography

Anand, K., Kirman, A., and Marsili, M. (2011). Epidemics of Rules, Rational Negligence and Market Crashes. *European Journal of Finance*, 0:1–10.

Anscombe, G. E. M. (1971). Causality and Determination. CUP Archive. Repr. in Sosa, E. and Tooley, M., editors, *Causation*, pages 88-104. Oxford: OUP, 1993.

Arthur, W. B. (2006). Out-of-equilibrium Economics and Agent-based Modeling. In Tesfatsion, L. and Judd, K. L., editors, *Handbook of Computational Economics. Agent-based Computational Economics*, volume 2, pages 1551–1564. North Holland: Elsevier.

Arthur, W. B., LeBaron, B., Palmer, B., and Taylor, R. (1997). Asset Pricing under Endogenous Expectations in an Artificial Stock Market. In Arthur, W. B., Durlauf, S. N., and Lane, D. A., editors, *Economy as an Evolving Complex System II*, volume XXVII, pages 15–44. Santa Fe Institute Studies in the Science of Complexity, Reading, MA: Addison-Wesley.

Atmanspacher, H. (2002). Determinism is Ontic, Determinability is Epistemic. In Atmanspacher, H. and Bishop, R., editors, *Between Chance and Choice*, pages 49–74. Thorverton: Imprint Academic.

Auyang, S. Y. (1998). *Foundations of Complex-systems Theories in Economics, Evolutionary Biology and Statistical Physics*. CUP.

Bak, P. (1997). *How Nature Works*. Oxford: OUP.

Baumgartner, M. (2008). Regularity Theories Reassessed. *Philosophia*, 36:327–354.

Bechtel, W. and Abrahamsen, A. (2005). Explanation: A Mechanist Alternative. *Studies in the History and Philosophy of the Biological and Biomedical Sciences*, 36:421–441.

Bechtel, W. and Richardson, R. C. (1992). Emergent Phenomena and Complex Systems. In Beckermann, A., Flohr, H., and Kim, J., editors, *Emer-*

gence or Reduction? Essays on the Prospects of Nonreductive Physicalism, pages 257–288. Berlin: Walter de Gruyter Verlag.

Bedau, M. A. (1997). Weak Emergence. *Philosophical Perspectives*, 11:375–399.

Bedau, M. A. (2002). Downward Causation and the Autonomy of Weak Emergence. *Principia*, 6(1):5–50.

Bedau, M. A. (2003). Artificial Life: Organization, Adaptation and Complexity From the Bottom Up. *Trends in Cognitive Sciences*, 7(11):505–512.

Bhalla, U. S. (2003). Understanding Complex Signaling Networks Through Models and Metaphors. *Progress in Biophysics & Molecular Biology*, 81:45–65.

Bhalla, U. S. and Iyengar, R. (1999). Emergent Properties of Networks of Biological Signaling Pathways. *Science*, 283:381–387.

Blackburn, S. (1990). Hume and Thick Connexions. *Philosophy and Phenomenological Research*, 50:237–250.

Blair, R. H., Kliebenstein, D. J., and Churchill, G. A. (2012). What Can Causal Networks Tell Us about Metabolic Pathways? *PLoS Computational Biology*, 8(4).

Block, N. (2003). Do Causal Powers Drain Away? *Philosophy and Phenomenological Research*, 67:133–150.

Brandom, R. B. (1994a). *Making It Explicit: Reasoning, Representing, and Discursive Commitment*. Harvard University Press.

Brandom, R. B. (1994b). Unsuccessful Semantics. *Analysis*, 54(3):175–178.

Brandom, R. B. (2000). *Articulating Reasons: An Introduction to Inferentialism*. Cambridge, MA: Harvard University Press.

Brandom, R. B. (2007). Inferentialism and Some of Its Challenges. *Philosophy and Phenomenological Research*, 74(3):651–676.

Brandom, R. B. (2008a). *Between Saying and Doing. Towards an Analytic Pragmatism*. Oxford: OUP.

Brandom, R. B. (2008b). Towards an Analytic Pragmatism. *Philosophical Topics*, 36(2):1–27.

Bruggeman, F. J. and Westerhoff, H. V. (2006). The Nature of Systems Biology. *Trends in Microbiology, 15(1): 45-50.*, 15(1):45–50.

Buchanan, M. (2009). Economics: Meltdown Modelling. could Agent-based Computer Models Prevent Another Financial Crisis? *Nature*, 460:680–682.

Bunge, M. (2004). How Does It Work? The Search for Explanatory Mechanisms. *Philosophy of the Social Sciences*, 34(1):182–210.

Campbell, R. (2009). A Process-based Model for an Interactive Ontology. *Synthese*, 166:453–477.

Carroll, J. W. (2009). Anti-Reductionism. In Beebee, H., Menzies, P., and Hitchcock, C., editors, *The Oxford Handbook of Causation*, pages 279–298. Oxford: OUP.

Cartwright, N. (1979). Causal Laws and Effective Strategies. *Nous*, 13:419–437.

Cartwright, N. (1989). *Nature's Capacities and Their Measurement*. Oxford: Clarendon Press.

Cartwright, N. (1994). Fundamentalism vs. the Patchwork of Laws. In *Proceedings of the Aristotelian Society*, volume 94, pages 279–292.

Cartwright, N. (1999). *The Dappled World: A Study of the Boundaries of Science*. Cambridge University Press.

Cartwright, N. (2004). Causation: One Word, Many Things. *Philosophy of Science*, 71:805–819.

Cartwright, N. (2007a). Causal Powers: What Are They? Why Do We Need Them? What Can Be Done with Them and What Cannot? Contingency and dissent in science, technical report 04/07, London School of Economics, Centre for Philosophy of Natural and Social Science.

Cartwright, N. (2007b). *Hunting Causes and Using Them*. Cambridge: CUP.

Cartwright, N. (2010). Comments on Longworth and Weber. *Analysis*, 70(2):325–330.

Cartwright, N. and Efstathiou, S. (2007). Hunting Causes and Using Them: Is There No Bridge from Here to There? Paper presented at the First Biennial conference of the Philosophy of Science in Practice, Twente University, August 2007.

Casini, L. (2012). Causation: Many Words, One Thing? *Theoria*, 27(74):203–219.

Casini, L., Illari, P., Russo, F., and Williamson, J. (2010). Recursive Bayesian Nets for Prediction, Explanation and Control in Cancer Science. In Fred, A., Filipe, J., and Gamboa, H., editors, *Proceedings of the First International Conference in Bioinformatics. INSTICC*.

Casini, L., Illari, P., Russo, F., and Williamson, J. (2011). Models for Prediction, Explanation and Control: Recursive Bayesian Networks. *Theoria*, 26(70):5–33.

Casti, J. L. (1994). *Complexification*. New York: Harper Collins.

Casti, J. L. (1997). *Would-be Worlds. How Simulation is Changing the Frontiers of Science*. New York: John Wiley & Sons.

Chadeau-Hyam, M., Athersuch, T. J., Keun, H. C., De Iorio, M., Ebbels, T. M., Jenab, M., Sacerdote, C., Bruce, S. J., Holmes, E., and Vineis, P. (2011). Meeting-in-the-middle Using Metabolic Profiling. A Strategy for the Identification of Intermediate Biomarkers in Cohort Studies. *Biomarkers*, 16(1):83–88.

Chakravartty, A. (2007). *A Metaphysics for Scientific Realism*. Cambridge University Press.

Chen, P. and Hsiao, C.-Y. (2010). Causal Inference for Structural Equations: With an Application to Wage-Price Spiral. *Computational Economics*, 36(1):17–36.

Chu, D. (2011). Complexity: Against Systems. *Theory in Biosciences*, 130(3):229–245.

Chu, D., Strand, R., and Fjelland, R. (2003). Theories of Complexity. *Complexity*, 8(3):19–30.

Cozzo, C. (2002). Does Epistemological Holism Lead to Meaning-Holism? *Topoi*, 21:25–45.

Craver, C. and Bechtel, W. (2007). Top-down Causation Without Top-down Causes. *Biology and Philosophy*, 22:547–563.

Dawid, H. and Neugart, M. (2011). Agent-Based Models for Economic Policy Design. *Eastern Economic Journal*, 37(1):44–50.

de Donato Rodríguez, X. and Zamora Bonilla, J. (2009a). Credibility, Idealisation, and Model Building: An Inferential Approach. *Erkenntnis*, 70(1):101–118.

de Donato Rodríguez, X. and Zamora Bonilla, J. (2009b). Explanation and Modelization in a Comprehensive Inferential Account. In *EPSA09: 2nd Conference of the European Philosophy of Science Association (Amsterdam, 21-24 October, 2009)*.

Descartes, R. ([1641] 1996). *Meditations on First Philosophy*. Cambridge: CUP. Trans. J. Cottingham.

deVries, W. A. (2010). Naturalism, the Autonomy of Reason, and Pictures. *International Journal of Philosophical Studies*, 18(3):395–413.

Dowe, P. (1995). Causality and Conserved Quantities: A Reply to Salmon. *Philosophy of Science*, 62:321–333.

Dummett, M. (1991). *The Logical Basis of Metaphysics*. Harvard University Press.

Dupré, J. (1993). *The Disorder of Things : Metaphysical Foundations of the Disunity of Science*. Cambridge, Mass: Harvard University Press.

Elga, A. (2000). Statistical Mechanics and the Asymmetry of Counterfactual Dependence. *Philosophy of Science*, 68:S313–S328.

Elster, J. (1989). *Nut and Bolts for the Social Sciences*. Cambridge University Press.

Elster, J. (1998). A Plea for Mechanisms. In Hedstrøm, P. and Swedberg, R., editors, *Social Mechanisms: An Analytical Approach to Social Theory*, pages 45–73. Cambridge University Press.

Emmeche, C., Køppe, S., and Stjernfelt, F. (2000). Levels, Emergence, and Three versions of Downward Causation. In Andersen, P. B., Emmeche, C., Finnemann, N. O., and Christiansen, P. V., editors, *Downward Causation. Minds, Bodies and Matter*, pages 13–34. Arhus: Aarhus University Press.

Epstein, J. M. (1999). Agent-based Computational Models and Generative Social Science. *Complexity*, 4(5):41–60.

Epstein, J. M. (2006). Remarks on the Foundations of Agent-based Generative Social Science. In Tesfatsion, L. and Judd, K. L., editors, *Handbook of Computational Economics. Agent-based Computational Economics*, volume 2, pages 1585–1604. North Holland: Elsevier.

Epstein, J. M. (2008). Why Model? *Journal of Artificial Societies and Social Simulation*, 11(4):12.

Érdi, P. (2008). *Complexity Explained*. Berlin Heidelberg: Springer Verlag.

Farmer, J. D. and Foley, D. (2009). The Economy Needs Agent-based Modelling. *Nature*, 460:685–686.

Fetzer, J. H. (1970). Dispositional Probabilities. In *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, pages 473–482.

Fetzer, J. H. (1981). *Scientific Knowledge. Causation, Explanation, and Corroboration*, volume 69 of *Boston Studies in the Philosophy of Science*. Dordrecht: D. Reidel.

Follmer, H., Horst, U., and Kirman, A. (2005). Equilibria in Financial Markets with Heterogeneous Agents: A Probabilistic Perspective. *Journal of Mathematical Economics*, 41(1-2):123–155.

Friedman, N., Linial, M., Nachman, I., and Pe'er, D. (2000). Using Bayesian networks to analyze expression data. *Journal of Computational Biology*, 7(3/4):601–620.

Frigg, R. (2003). Self-organised Criticality—What It Is and What It Isn't. *Studies in History and Philosophy of Science*, 34:613–632.

Frigg, R. (2006). Scientific Representation and the Semantic View of Theories. *Theoria*, 55:37–53.

Frigg, R. and Reiss, J. (2009). The philosophy of Simulation: Hot New Issues or Same Old Stew? *Synthese*, 169(3):593–613.

Frisch, M. (2012). No Place for Causes? Causal Skepticism in Physics. *European Journal of Philosophy of Science*, 2(3):313–336.

Galea, S., Riddle, M., and Kaplan, G. A. (2010). Causal Thinking and Complex System Approaches in Epidemiology. *International Journal of Epidemiology*, 39:97–106.

Gersherson, C. (2002). Complex Philosophy. URL = http://arxiv.org/ftp/nlin/papers/0109/0109001.pdf.

Giere, R. (1988). *Explaining Science: A Cognitive Approach*. Chicago: University of Chicago Press.

Giere, R. (2003). Perspectival Pluralism. URL = www.tc.umn.edu/~giere/pp.pdf.

Giere, R. N. (2005). Scientific Realism: Old and New Problems. *Erkenntnis*, 63:149–165.

Gilbert, N. and Troitzsch, K. (2005). *Simulation for the Social Scientist*. Open University Press, 2nd edition.

Glennan, S. (1996). Mechanisms and the Nature of Causation. *Erkenntnis*, 44:49–71.

Glennan, S. (1997). Capacities, Universality and Singularity. *Philosophy of Science*, 64:605–626.

Glennan, S. (2000). A Model of Models. URL = http://philsci-archive.pitt.edu/1134/.

Glennan, S. (2002). Rethinking Mechanistic Explanation. *Philosophy of Science. Supplement: Proceedings of the 2000 Biennial Meeting of the Philosophy of Science Association. Part II: Symposia Papers (Sep., 2002)*, 69(3):S342–S353.

Glennan, S. (2005). Modeling Mechanisms. *Studies in History and Philosophy of Biology & Biomedical Sciences*, 36:443–464.

Glennan, S. (2008). Mechanisms. In Psillos, S. and Curd, M., editors, *The Routledge Companion to the Philosophy of Science*, pages 376–384. London: Routledge.

Glennan, S. (2010). Mechanisms, Causes, and the Layered Model of the World. *Philosophy and Phenomenological Research*, 81(2):362–381.

Glennan, S. (2011). Singular and General Causal Relations: A Mechanist Perspective. In Illari, P., Russo, F., and Williamson, J., editors, *Causality in the Sciences*. Oxford: OUP.

Godfrey-Smith, P. (2009). Causal Pluralism. In Beebee, H., Menzies, P., and Hitchcock, C., editors, *The Oxford Handbook of Causation*, pages 326–337. Oxford: OUP.

Goldstein, J. (1996). Causality and Emergence in Chaos and Complexity Theories. In Sulis, W. H. and Combs, A., editors, *Nonlinear Dynamics in Human Behavior*, pages 161–190. World Scientific Publishing.

Granger, C. W. J. (1969). Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica*, 37(3):424–438.

Guala, F. (2012). Experimentation in Economics. In Mäki, U., editor, *Philosophy of Economics*, volume 13 of *Handbook of the Philosophy of Science.*, pages 597–640. Elsevier.

Haavelmo, T. (1943). The Statistical Implications of a System of Simultaneous Equations. *Econometrica*, 11:1–12.

Hall, N. (2004). Two Concepts of Causation. In Collins, J., Hall, N., and Paul, L. A., editors, *Causation and Counterfactuals*, pages 225–276. Cambridge, MA: MIT Press.

Harman, G. (1999). *Reasoning, Meaning and Mind.* Oxford University Press.

Hartwell, L. H., Hopfield, J. J., Leibler, S., and Murray, A. W. (1999). From Molecular to Modular Cell Biology. *Nature*, 402:C47–C52.

Hedstrøm, P. and Swedberg, R. (1998). Social Mechanisms: An Introductory Essay. In Hedstrøm, P. and Swedberg, R., editors, *Social Mechanisms: An Analytical Approach to Social Theory*, pages 1–31. Cambridge University Press.

Hesslow, G. (1976). Discussion: Two Notes on the Probabilistic Approach to Causality. *Philosophy of Science*, 43:290–292.

Heylighen, F. (2001). The Science of Self-organization and Adaptivity. In Kiel, L. D., editor, *Knowledge Management, Organizational Intelligence and Learning, and Complexity, in: The Encyclopedia of Life Support Systems (EOLSS).* Oxford: Eolss Publishers.

Hiddleston, E. (2005). *The Philosophical Review*, 114(4):545–547. Review of *Making Things Happen* by J. Woodward.

Hitchcock, C. (2001). The Intransitivity of Causation Revealed in Equations and Graphs. *Journal of Philosophy*, 98:273–299.

Hitchcock, C. (2007). How To Be a Causal Pluralist. In Machamer, P. K. and Wolters, G., editors, *Thinking About Causes*, pages 200–221. Pittsburgh: University of Pittsburgh Press.

Hitchcock, C. (2010). Probabilistic Causation. E. N. Zalta (ed.): *The Stanford Encyclopedia of Philosophy (Fall 2010 Edition)*. URL = http://plato.stanford.edu/archives/fall2010/entries/causation-probabilistic/.

Holland, J. (1995). *Hidden Order: How Adaptation Builds Complexity*. Reading, MA: Addison-Wesley.

Holland, P. W. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association*, 81(396):945–960.

Hommes, C. H. (2006). Heterogeneous Agent Models in Economics and Finance. In Tesfatsion, L. and Judd, K. L., editors, *Handbook of Computational Economics. Agent-based Computational Economics*, volume 2, pages 1109–1186. North Holland: Elsevier.

Hornberg, J. J., Bruggeman, F. J., Westerhoff, H. V., and Lankelma, J. (2006). Cancer: A Systems Biology Disease. *Biosystems*, 83:81–90.

Hume, D. ([1740] 1968). *A Treatise of Human Nature*. Oxford: Clarendon Press, 1888 edition.

Hume, D. ([1748] 1975). *An Enquiry Concerning Human Understanding*. Oxford: Clarendon Press, 1777 edition.

Humphreys, P. (2008a). Computational and Conceptual Emergence. *Philosophy of Science*, 75(5):S584–S594.

Humphreys, P. (2008b). Synchronic and Diachronic Emergence. *Minds & Machines*, 18:431–442.

Humphreys, P. (2009). The Philosophical Novelty of Computer Simulation Methods. *Synthese*, 169(3):615–626.

Humphreys, P. W. (1997a). Emergence, Not Supervenience. *Philosophy of Science*, 64(4):S337–S345.

Humphreys, P. W. (1997b). How Properties Emerge. *Philosophy of Science*, 64(1):1–17.

Israel, G. (2005). The Science of Complexity: Epistemological Problems and Perspectives. *Science in Context*, 18(3):1–31.

Kauffman, S. (1993). *Origins of Order: Self-Organization and Selection in Evolution*. Oxford: OUP.

Kim, J. (1999). Making Sense of Emergence. *Philosophical Studies*, 95(1-2):3–36.

Kim, J. (2005). *Physicalism, or Something Near Enough*. Princeton University Press.

Kineman, J. J. (2011). Relational Science: A Synthesis. *Axiomathes*, 21:393–437.

King, R. J. B. (2000). *Cancer Biology*. Pearson, Singapore, 2nd edition.

Kirman, A. (2010). The Economic Crisis is a Crisis for Economic Theory. *CESifo Economic Studies*, 56(4):498–535.

Kitano, H. (2002a). Computational Systems Biology. *Nature*, 420:206–210.

Kitano, H. (2002b). Systems Biology: a Brief Overview. *Science*, 295(5560):1662–1664.

Klipp, E., Liebermeister, W., Wierling, C., Kowald, A., Lehrach, H., and Herwig, R. (2009). *Systems Biology: A Textbook*. Weinheim: Wiley-VCH.

Korb, K. B. and Nyberg, E. (2006). The Power of Intervention 16:289-302. *Mind and Machines*, 16:289–302.

Kuhlmann, M. (2011). Mechanisms in Dynamically Complex Systems. In Illari, P., Russo, F., and Williamson, J., editors, *Causality in the Sciences*, pages 880–906. Oxford: OUP.

Kuipers, B. (1987). Qualitative Simulation as Causal Explanation. In *IEEE Transactions on Systems, Man, and Cybernetics*, volume 17, pages 432–444.

Laplace, P. S. ([1814] 1902). *A Philosophical Essay on Probabilities*. New York: John Wiley & Sons. Trans. F. W. Truscott and F. L. Emory.

Laurence, S. and Margolis, E. (1999). Concepts and Cognitive Science. In Laurence, S. and Margolis, E., editors, *Concepts: Core Readings*, pages 3–81. Cambridge, MA: MIT Press.

Lazebnik, Y. (2002). Can a Biologist Fix a Radio?—Or, What I Learned While Studying Apoptosis. *Cancer Cell*, 2:179–182.

LeBaron, B. (2002). Building the Santa Fe Artificial Stock Market. Working Paper, Brandeis University, June 2002.

LeBaron, B. (2006). Agent-based Computational Finance. In Tesfatsion, L. and Judd, K. L., editors, *Handbook of Computational Economics. Agent-based Computational Economics*, volume 2, pages 1187–1233. North Holland: Elsevier.

LeBaron, B. D., Arthur, W. B., and Palmer, R. G. (1999). Time Series Properties of an Artificial Stock Market. *Journal of Economic Dynamics and Control*, 23:1487–1516.

Legewie, S., Blüthgen, N., and Herzel, H. (2006). Mathematical Modeling Identifies Inhibitors of Apoptosis as Mediators of Positive Feedback and Bistability. *PLoS Computational Biology*, 2(9):1061–1073.

Lewis, D. (1986). *Philosophical Papers*, volume II. New York: Oxford University Press.

Lewis, D. (2004). Causation as Influence. In Collins, J., Hall, N., and Paul, L. A., editors, *Causation and Counterfactuals*, pages 75–106. Cambridge, MA: MIT Press.

Little, D. (2006). Levels of the Social. In Risjord, M. and Turner, S., editors, *The Philosophy of Anthropology and Sociology*, pages 343–371. Elsevier Science.

Longworth, F. (2006). Causation, Pluralism and Moral Responsibility. *Philosophica*, 77(1):45–68.

Longworth, F. (2010). Cartwright's Causal Pluralism: a Critique and an Alternative. *Analysis*, 70(2):310–318.

Lorenz, E. (1993). *The Essence of Chaos*. London: UCL Press.

Louie, A. H. (2010). Robert Rosen's Anticipatory Systems. *Foresight*, 12(3):18–29.

Lux, T. and Marchesi, M. (1999). Scaling and Criticality in a Stochastic Multi-agent Model of a Financial Market. *Nature*, 397:498–500.

Lux, T. and Marchesi, M. (2000). Volatility Clustering in Financial Markets: A Microsimulation of Iinteractive Agents. *International Journal of Theoretical and Applied Finance*, 3(4):675–702.

Macal, C. M. and North, M. J. (2005). Tutorial on Agent-based Modeling and Simulation. In Kuhl, M. E., Steiger, N. M., Armstrong, F. B., and Joines, J. A., editors, *Proceedings of the 2005 Winter Simulation Conference*, pages 2–15.

Machamer, P., Darden, L., and Craver, C. (2000). Thinking about Mechanisms. *Philosophy of Science*, 67:1–25.

Mackie, J. L. (1974). *The Cement of the Universe. A Study of Causation.* Oxford: Clarendon Press.

Mai, Z. and Liu, H. (2009). Boolean Network-based Analysis of the Apoptotis Network: Irreversible Apoptosis and Stable Surviving. *Journal of Theoretical Biology*, 259:760–769.

Margolis, E. and Laurence, S. (2012). Concepts. E. N. Zalta (ed.): *The Stanford Encyclopedia of Philosophy (Fall 2012 Edition)*. URL = http://plato.stanford.edu/archives/fall2012/entries/concepts.

Markose, S., Arifovic, J., and Sunder, S. (2007). Advances in Experimental and Agent-based Modelling: Asset Markets, Economic Network, Computational Mechanism Design and Evolutionary Game Dynamics. *Journal of Economic Dynamics and Control*, 31:1801–1807.

Materi, W. and Wishart, D. S. (2007). Computational Systems Biology in Drug Discovery and Development: Methods and Applications. *Drug Discovery Today*, 12(7/8):295–303.

Menzies, P. (1996). Probabilistic Causation and the Pre-emption Problem. *Mind*, 105:85–117.

Menzies, P. (2009). Counterfactual Theories of Causation. E. N. Zalta (ed.): *The Stanford Encyclopedia of Philosophy (Fall 2009 Edition)*. URL = http://plato.stanford.edu/archives/fall2009/entries/causation-counterfactual/.

Menzies, P. and Price, H. (1993). Causation as a Secondary Quality. *British Journal for the Philosophy of Science*, 44:187–203.

Mikulecky, D. C. (2001). The Emergence of Complexity: Science Coming of Age or Science Growing Old? *Computers and Chemistry*, 25:341–348.

Mikulecky, D. C. (2007). Causality and Complexity: The Myth of Objectivity in Science. *Chemistry and Biodiversity*, 4:2480–2491.

Mitchell, S. D. (2003). *Biological Complexity and Integrative Pluralism*. Cambridge: CUP.

Morgan, M. and Morrison, M., editors (1999). *Models as Mediators*. Cambridge: CUP.

Newman, M., Barabási, A.-L., and Watts, D. J. (2006). *The Structure and Dynamics of Networks*. Princeton, NJ: Princeton University Press.

Nicolis, G. and Prigogine, I. (1989). *Exploring Complexity*. New York: W. H. Freeman and Company.

Norton, J. D. (2003). Causation as Folk Science. *Philosopher's Imprint*. URL = http://www.philosophersimprint.org/003004/.

O'Connor, T. and Wong, H. Y. (2005). The Metaphysics of Emergence. *Noûs*, 39:658–678.

O'Connor, T. and Wong, H. Y. (2009). Emergent Properties. E. N. Zalta (ed.): *The Stanford Encyclopedia of Philosophy*. URL = http://plato.stanford.edu/archives/spr2008/entries/properties-emergent/.

O'Malley, M. A. and Dupré, J. (2005). Fundamental Issues in Systems Biology. *BioEssays*, 27:1270–1276.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo, CA: Morgan Kaufmann.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge.

Peregrin, J. (2008). Brandom's Incompatibility Semantics. *Philosophical Topics*, 36(2):99–121.

Peregrin, J. (2012). Inferentialism and the Normativity of Meaning. *Philosophia*, 40:75–97.

Price, H. (1998). Two Paths to Pragmatism II. *European Journal of Philosophy*, 3.

Prigogine, I. and Stengers, I. (1984). *Order Out of Chaos*. Glasgow: Flamingo.

Psillos, S. (2004). A Glimpse of the Secret Connexion: Harmonising Mechanisms with Counterfactuals. *Perspectives on Science*, 12(3):288–319.

Psillos, S. (2010). Causal Pluralism. In Vanderbeeken, R. and D'Hooghe, B., editors, *Worldviews, Science and Us: Studies of Analytical Metaphysics. A Selection of Topics from a Methodological Perspective*, pages 131–151. Singapore: World Scientific Publishers.

Putnam, H. (2002). *The Collapse of the Fact/value Dichotomy and Other Essays*. Harvard University Press.

Quine, W. V. O. (1951). Two Dogmas of Empiricism. *Philosophical Review*, 60(1):20–43. Reprinted in *From a Logical Point of View*, 2nd ed., 1961, New York and Evanston: Harper & Row, pp. 20-46.

Reichenbach, H. (1956). *The Direction of Time*. Berkeley and Los Angeles: University of California Press, 1971 edition.

Reiss, J. (2007). Time Series, Nonsense Correlatations and the Principle of the Common Cause. In *Causality and Probability in the Sciences*, pages 179–196. London: College Publications.

Reiss, J. (2009a). Causation in the Social Sciences. Evidence, Inference, and Purpose. *Philosophy of the Social Sciences*, 39(1):20–40.

Reiss, J. (2009b). Causation, Inference and Wittgenstein. Unpublished manuscript.

Reiss, J. (2011). Third Time's a Charm: Causation, Science and Wittgensteinian Pluralism. In Illari, P., Russo, F., and Williamson, J., editors, *Causality in the Sciences*, pages 907–927. Oxford: OUP.

Reiss, J. (2012). Causation in the Sciences: An Inferentialist Account. *Studies in History and Philosophy of Biological and Biomedical Sciences*, 43(4):769–777.

Rickles, D. (2009). Causality in Complex Interventions. *Medicine, Health Care and Philosophy*, 12(1):77–90.

Rickles, D. (2011). Econophysics and the Complexity of Financial Markets. In Collier, J. and Hooker, C., editors, *Philosophy of Complex Systems*, volume 10 of *Handbook of the Philosophy of Science.*, pages 527–561. North Holland: Elsevier.

Rosen, R. (1985). *Anticipatory Systems.* New York: Pergamon Press.

Rosen, R. (1991). *Life Itself.* New York: Columbia University Press.

Rosen, R. (1998). *Essays on Life Itself.* Columbia University Press.

Ross, S. M. (2003). *An Elementary Introduction to Mathematical Finance. Options and Other Topics.* Cambridge University Press, 2nd edition.

Russell, B. (1913). On The Notion of Cause. *Proceedings of the Aristotelian Society*, 13:1–26.

Russo, F. and Williamson, J. (2007). Interpreting Causality in the Health Sciences. *International Studies in the Philosophy of Science*, 21(2):157–170.

Salmon, W. (1997). Causality and Explanation: A Reply to Two Critiques. *Philosophy of Science*, 61:461–477.

Samanidou, E., Zschischang, E., Stauffer, D., and Lux, T. (2007). Agent-based Models of Financial Markets. *Reports on Progress in Physics*, 70:409–450.

Sawyer, R. K. (2004). The Mechanisms of Emergence. *Philosophy of the Social Sciences*, 34(2):260–282.

Schaffer, J. (2000). Causation by Disconnection. *Philosophy of Science*, 67:285–300.

Sellars, W. (1948). Concepts as Involving Laws, and Inconceivable Without Them. *Philosophy of Science*, 15(4):287–315.

Sellars, W. (1949). Language, Rules and Behavior. In Hook, S., editor, *John Dewey: Philosopher of Science and Freedom*, pages 289–315. New York: Dial Press. Reprinted in: J. F. Sicha (ed.), *Pure Pragmatics and Possible Worlds. The Early Essays of Wilfrid Sellars*, Atascadero, CA: Ridgeview Publishing Co, 1980, pages 129-155.

Sellars, W. (1953). Inference and Meaning. *Mind*, 62(247):313–338.

Sellars, W. (1962). Truth and Correspondence. *The Journal of Philosophy*, 59(2):29–56.

Sellars, W. (1968). *Science and Metaphysics: Variations on Kantian Themes*. London: Routledge & Kegan Paul.

Silberstein, M. and McGeever, J. (1999). The Search for Ontological Emergence. *The Philosophical Quarterly*, 49:182–200.

Simon, H. A. (1996). *The Sciences of the Artificial*. MIT Press, 3rd edition.

Simon, H. A. (2000). Bounded Rationality in Social Science: Today and Tomorrow. *Mind and Society*, 1(1):25–39.

Skovgaard Olsen, N. (2012). Brandom, TCA, and the Social Foundation of Objectivity. Submitted to the *International Journal of Philosophical Studies*.

Smith, P. (1998). *Explaining Chaos*. Cambridge: CUP.

Sornette, D. (2002). Predictability of Catastrophic Events: Material Rupture, Earthquakes, Turbolence, Financial Crashes, and Human Birth. *Proceedings of the National Academy of Sciences of the United States of America*, 99(1):2522–2529.

Soros, G. (1987). *The Alchemy of Finance. Reading the Mind of the Market*. New York: Simon and Schuster.

Spirtes, P., Glymour, C., and Scheines, R. (1993). *Causation, Prediction, and Search*. MIT Press, Cambridge MA, second (2000) edition.

Spohn, W. (2002). Bayesian Nets Are All There Is to Causal Dependence. In Galavotti, M. C., Suppes, P., and Costantini, D., editors, *Stochastic Causality*. Chicago, IL: University of Chicago Press.

Steel, D. (2004). Social Mechanisms and Causal Inference. *Philosophy of the Social Sciences*, 34(1):55–78.

Steel, D. (2007). *Across the Boundaries*. Oxford: OUP.

Stewart, I. (1997). *Does God Play Dice?* London: Penguin Books, 2nd edition.

Strawson, G. (1989). *The Secret Connexion. Causation, Realism, and David Hume*. Oxford: OUP.

Strevens, M. (2007). Review of Woodward, *Making Things Happen. Philosophy and Phenomenological Research*, 74(1):233–249.

Strogatz, S. H. (1994). *Nonlinear Dynamics and Chaos.* Reading, MA: Perseus Books.

Suárez, M. (2003). Scientific Representation: Against Similarity and Isomorphism. *International Studies in the Philosophy of Science*, 17:225–244.

Suárez, M. (2004). An Inferential Conception of Scientific Representation. *Philosophy of Science*, 71(5):767–779.

Suppe, F. (1989). *The Semantic View of Theories and Scientific Realism.* Urbana and Chicago: University of Illinois Press.

Tesfatsion, L. (2002). Agent-based Computational Economics: Growing Economies From the Bottom Up. *Artificial Life*, 8(1):55–82.

Tesfatsion, L. (2006). Agent-based Computational Economics: A Constructive Approach to Economic Theory. In Tesfatsion, L. and Judd, K. L., editors, *Handbook of Computational Economics. Agent-based Computational Economics*, volume 2, pages 831–880. North Holland: Elsevier.

Thurner, S., Farmer, J. D., and Geanakoplos, J. (2010). Leverage Causes Fat Tails and Clustered Volatility. Cowles Foundation Discussion Papers 1745, Cowles Foundation for Research in Economics, Yale University.

Twardy, C. R. and Korb, K. B. (2004). A Criterion of Probabilistic Causality. *Philosophy of Science*, 71:241–262.

van Fraassen, B. (1980). *The Scientific Image.* Oxford: OUP.

Varela, F., Maturana, H., and Uribe, G. (1974). Autopoiesis: the Organisation of Living Systems. Its Characterisation and a Model. *Biosystems*, 5:187–96.

Vineis, P., Khan, A., Vlaanderen, J., and Vermeulen, R. (2009). The Impact of New Research Technologies on Our Understanding of Environmental Causes of Disease: the Concept of Clinical Vulnerability. *Environmental Health*, 8(1):54.

Wagner, A. (1999). Causality in complex systems. *Biology and Philosophy*, 14:83–101.

Weinberg, R. A. (2007). *The Biology of Cancer*. Garland Science, Taylor & Francis Group, New York.

Weintraub, E. R. (1993). Neoclassical Economics. In *The Concise Encyclopedia of Economics*. Library of Economics and Liberty. URL = http://www.econlib.org/library/Enc1/NeoclassicalEconomics.html.

Weisberg, M. (2005). Water is not H_2O. In Baird, D., Scerri, E., and MacIntyre, L., editors, *Philosophy of Chemistry: Synthesis of a New Discipline*, pages 337–345. New York: Springer.

Weng, G., Bhalla, U. S., and Iyengar, R. (1999). Complexity in Biological Signaling Systems. *Science*, 284:92–96.

Westerhoff, H. V. and Kell, D. B. (2007). The Methodologies of Systems Biology. In Boogerd, F. C., Bruggeman, F. J., Hofmeyr, J. S., and Westerhoff, H. V., editors, *Systems Biology. Philosophical Foundations*, pages 23–70. North Holland: Elsevier.

Williamson, J. (2005). *Bayesian Nets and Causality: Philosophical and Computational Foundations*. Oxford: OUP.

Williamson, J. (2006). Causal Pluralism versus Epistemic Causality. *Philosophica*, 77(1):69–96.

Wimsatt, W. (2007). *Re-Engineering Philosophy for Limited Beings*. Cambridge, MA: Harvard University Press.

Winsberg, E. (1999). Sanctioning Models: The Epistemology of Simulation. *Science in Context*, 12(2):275–292.

Winsberg, E. (2009). A Tale of Two Methods. *Synthese*, 169(3):575–592.

Wittgenstein, L. ([1956] 1978). *Bemerkungen über die Grundlagen der Mathematik*. Oxford: Basil Blackwell. English Trans. *Remarks on the Foundations of Mathematics*.

Wolkenhauer, O. (2001). Systems Biology: The Reincarnation of Systems Theory Applied in Biology? *Briefings in Bioinformatics*, 2(3):258–270.

Wolkenhauer, O. and Ullah, M. (2007). All Models Are Wrong...Some More Than Others. In Boogerd, F. C., Bruggeman, F. J., Hofmeyr, J. S., and Westerhoff, H. V., editors, *Systems Biology. Philosophical Foundations*, pages 163–180. North Holland: Elsevier.

Woodward, J. (2002). What is a Mechanism? A Counterfactual Account. *Philosophy of Science*, 69:S366–S377.

Woodward, J. (2003). *Making Things Happen. A Theory of Causal Explanation*. New York: Oxford University Press.

Woodward, J. (2008). Causation and Manipulability. E. N. Zalta (ed.): *The Stanford Encyclopedia of Philosophy (Winter 2008 Edition)*. URL = http://plato.stanford.edu/archives/win2008/entries/causation-mani/.

Yoo, C., Thorsson, V., and Cooper, G. F. (2002). Discovery of Causal Relationships in a Gene-regulation Pathway from a Mixture of Experimental and Observational DNA Microarray Data. In *Pacific Symposium on Biocomputing*, volume 7, pages 498–509.

Zamora Bonilla, J. (2006). Science as a Persuasion Game. An Inferentialist Approach. *Episteme*, 2:189–201.