Nathalie Bulle (2009), Under What Conditions Can Formal Models of Social Action Claim Explanatory Power? *International Studies in the Philosophy of Science*, 23: 47-64.

Under What Conditions Can Formal Models of Social Action Claim Explanatory Power?

Abstract

This paper's purpose is to set forth the conditions of explanation in the domain of formal modelling of social action. Explanation is defined as an adequate account of the underlying factors bringing about a phenomenon. The modelling of a social phenomenon can claim explanatory value in this sense if the following two conditions are fulfilled. (1) The generative mechanisms involved translate the effects of real factors abstracted from their phenomenal context, not those of purely ideal ones. (2) The explanatory hypotheses, which account for the effects of explanatory factors, and the purely descriptive hypotheses, which introduce conceptual simplifications and summarize complex secondary mechanisms, are relatively independent from each other with regard to the phenomenon represented. This condition subjects the model to testing by alternatives through the development of purely descriptive hypotheses in the sense of explanatory or analytical realism.

1. Introduction

A growing interest in social sciences research focuses on explication, claiming that looking for the generative mechanisms underlying social phenomena is of primary importance for research.[1] A generative mechanism can be defined as an intellectual representation of the specific combination of factors that genuinely brings about a given social phenomenon. The idea of "generative mechanism" is rooted in the differentiation of theoretical levels of explanandum and explanans as emphasized on pluridisciplinary grounds by Stinchcombe (1991) and Manicas (2006). In other words, such mechanisms typically deal in the social and human sciences with actors and their situations. In this theoretical context, the formal modelling of social action represents an essential tool of scientific investigation. It allows us to implement and test the details of logical links between macro situations, interactions between agents, aggregation processes and emergent social properties. By “emergent” we mean that such social properties are the result of the interdependence of individuals’ actions and therefore cannot be deduced from individual units’ properties.

Nevertheless, with respect to the modelling of social mechanisms, characterizing explanation as opposed to description is no more obvious than in other scientific domains. It depends heavily upon the particular approaches to the general problem of the limitation of formalism. A scientific model places two worlds in relation, neither of which can be reduced to the other: a closed, formal world and a real, open one. The operations of abstraction and formalization involved in the model-making process introduce an irreducible distance between the phenomenon and its representation. The formal explanatory factors only retain certain aspects of the phenomenon under consideration; and the model is formally autonomous in relation to reality. The factors it brings into play are operationally defined by the relations it focuses on, and only by these relations- i.e. by the whole set of relations linking the factors of the model. They are therefore not simply abstracted from the whole set of factors that are operating in reality, but are literally *reconstructed*. The possibility of « manipulating » the model in order to deduce mechanically some group of consequences in the phenomenon under study depends on this relative closed nature of the set of concepts and relations from which the model is constructed. But the capacity of the model to account for this phenomenon in terms of the mechanisms that genuinely produce it is problematic. The links between factors that operate in reality and formal factors are usually envisioned only by means of analogical relationships between theoretical model and represented phenomenon. In some cases models are thought of as “highly abstract thought experiment(s) that explore plausible mechanisms that may underlie observed patterns” (Macy and Willer, 2002 :147).

The under-theorization of the relations between model and phenomenon leads formal modelling to be centered on the condition of « generative sufficiency » (Epstein & Axtel 1996 ; Epstein 2007 ; Hedström 2005 :143-4) or on the pattern of « inductive inference » (Sugden :2000).[2] This approach to explanation is shared by authors in the field whose works are among the most advanced in the rapidly increasing literature on agent-based modelling.[3] However, analogical relationships between explanatory formal factors in a model and real factors, and between properties generated and phenomenon under study, still do not allow us to distinguish a mere interpretation from an explanation, properly so called.

With regard to a potential transformation of major modes of formal analysis in sociology, we aim to lay out the general conditions for formal models of social action to possess explanatory power, i.e. conditions for an adequate account of the underlying factors bringing about a social phenomenon. Our claim is that two general conditions have to be respected, the existence of real correlates for formal explanatory factors and the validity of postulated epistemic correlations between formal and real factors. Let us note that in terms of what is applied to the world of experience, "real" means "with a traceable effect" for us. We conclude that the explanatory scope of models is fundamentally a matter of analytic relevance, as opposed to empirical adequacy with regard to the generative mechanisms represented. What counts, from this point of view, are the explanatory sub-structures of the models, that which models entail with regard to the real factors involved. According to our analysis, the conditions we have defined match up with certain central principles developed in the philosophy of science, but form part of a more general problem-set, in which there may be no possibility of isolating (not even in theory) the real factors involved.

2. Explanation with Regard to Formal Models

In the broadest sense a model is an « interpretive description » (Bailer-Jones 2003 : 61). Its ability to represent a specific phenomenon depends on an interpretation that goes beyond pure phenomenological reality. In this respect we also refer to *explanation.* But one may distinguish two modes of explanation in science, explanation in the weak sense and explanation in the strong sense. The distinction of two forms of scientific explanation appears in various guises in the literature, regarding the natural and physical sciences as well as the social sciences.[4] The first consists in subsuming the phenomenon to be represented under a general theoretical frame, in order to transcribe its behavior. The second consists in describing it in terms of the forces or factors that genuinely produce it. In the first sense different potential formal models are supposed to be able to describe the same phenomenon as a function of the instrumental objectives in view. In sense 2 one aims a priori at a single explanatory theory, without necessarily expecting the theory to be the last word on the matter.

Attempts at modelling the underlying factors that generate a phenomenon X presuppose that a theoretical point of view has been adopted and that the particular goal being pursued is related to criteria of truth or falsity. Note that on this point nor the argument referring to redundancy used by Pierre Duhem, nor the argument referring to theoretical invention used by Bas van Fraassen (1980) against inference to the best explanation (explanation by subsuming under a theory), affects causal explanation.[5] As stated by Nancy Cartwright (1983: 87-99), inference from effect to cause is legitimate. Such an inference is based on the existence of the cause under consideration, and the question of this existence is immune to any problem of redundancy (see also Suarez 2008).

Let us examine the point of view of van Fraassen. This epistemologist subscribes to a semantic conception of theories according to which they represent sets of models, considered as classes of structures. What counts for the epistemologist are the "empirical sub-structures" of models: that which the models imply with regard to that which is observable. These sub-structures can be compared to the sub-structures of phenomena or models of data. Explanation does not send us back to any theoretical truth, but in essence refers us to empirical adequacy, that is, to the development of models whose observable sub-structures are isomorphic with the data from experience.

Van Fraassen is in error about the claims of explanation in the strong sense. He considers it cannot be substantially true, although it only claims to express the truth of factors involved, that is, to give an adequate account of their role in the production of a given phenomenon. What is at stake is in reality the explanatory sub-structure of models, which we call their theoretical core. On this subject, as Ian Hacking emphasizes (1983:77), we can for example measure IQ with a dozen different techniques and even prove that they all provide the same result; we still have not produced any part of a causal explanation of the phenomenon. Concepts of modelling as based upon the construction of isomorphisms, partial isomorphisms,[6] or even upon simple relations of similarity[7] between model and phenomenon, thus lack an ingredient central to explanation: the existence of real correlates of formal explanatory factors- i.e. factors whose effects underlie the production of the phenomenon in the model. However, the search for truth does not concern the nature of the real correlates as such, but only the effects that are attributed to them in the framework of the model.

This existence of real correlates of formal explanatory factors is problematic, nonetheless, if following Cartwright we consider as suspect the composition of causes. Theories are only true of the objects of models, and the objects of the model are “simulacra” of objects in reality: they have only ‘the form or appearance of things’ and, in a very strong sense, not their ‘substance or proper qualities’ (Cartwright 1983:17). The possibility of combining the action of factors of different kinds, that is, on the basis of our knowledge of their behavior in different domains, is cast into doubt. That explains the interest shown by Cartwright in the operating role of factors or singular processes which she calls “capacities”, referring to the notion of tendency[8] derived from Mill and identifying “those tendencies which are tendencies to cause or to bring about something”. Capacities designate causal tendencies that continue to produce their effects, in various situations, by interfering with the course of other factors or processes. Their generality, given a sufficient degree of abstraction, distinguishes them from causal imputations subject to *ceteris paribus* conditions (requiring in an exorbitant fashion, in order to be applicable, non-interference with other causal factors).

But one cannot escape the co-implication of mechanisms involved in the social sciences. For the partisans of methodological individualism, the rationality of social actors represents an elementary trans-situational consistency. It offers an illustration of the notion of “capacity”, and represents a gauge of the explanatory potential of models, but is far from sufficient in order to justify their pertinence from an explanatory point of view. Although it represents a tendency or fundamental human capacity, it cannot be defined operationally in a trans-situational manner. It comes out of processes that are necessarily subjective or “limited”. To express the impossibility of accounting for human rationality *ab abstracto*, we qualify it more accurately as “cognitive” (Boudon 2003).[9]

Cartwright insists upon the tenuous nature, in regard to reality, of explanations in the strong sense: “truth doesn’t explain much” (Cartwright 1983:44-54). But we are nonetheless tempted to suppose that explanation may cover a larger field, if we make it dependent on the accuracy of the roles played by explanatory factors in the framework of models - i.e. on the attribution, via the model, of the right effect to the right real factor involved. In this sense explanation and description each take on a relative status. Explanation unveils specific combination of factors involved in generative mechanisms whereas description just sums up the joint effects of these factors. Theoretical analysis may pursue involving factors of a greater generality. The greatest generality refers back to a trans-situational truth in the sense of the concept of capacity in Cartwright.

This approach to explanation fits in with a propositional approach to representation (Bailer-Jones 2003: 60) following which models can be expressed in terms of propositions about the empirical world. When a model represents a phenomenon, the model entails propositions that are true with regard to that phenomenon. From this we deduce that a question about explanation- i.e. about factors that genuinely brings about that phenomenon- applies to the possibility of deriving from the model, that is, from the analysis of a closed system, a set of true propositions concerning the production of a given phenomenon within the open system of the phenomenal world. The truth in question is a sort of limit concept as regards reasonable belief: it refers us to what is ultimately justifiable (Ellis 1996:187).

Let the phenomenon X represent an underlying regularity or tendency, that is, one which independently give rise to theoretical reasoning and may subsume a set of social realities of greater or lesser importance:[10] a tendency toward segregation in certain urban areas, a tendency for used car markets to grow to a sub-optimal size, a tendency for social mobility to remain stable despite the expansion of educational systems, etc. As stated above, what we shall term the explanatory core of a model that purports to account for phenomenon X, is the theoretical substructure of the model: a theory Y which the formal model represents; and what we shall consider as explanatory power applies to the derivation from the model of a set of true propositions concerning the production of X. Our problem can now be reformulated in these terms*: to what extent can the theory Y claim to provide support for true propositions about the underlying factors bringing about phenomenon X* ?

3. Epistemic Correlations Are What Allow Us to Speak of Truth

Filmer Northrop’s (1947) theory of concepts as well as Henry Margenau’s theory of knowledge (1935, 1950), two approaches that are consonant with each other, may help to provide some clarification here. Northrop distinguishes “concepts by intuition” and “concepts by inspection” from concepts by postulation, or scientific concepts. Concepts by intuition or by inspection have “denotation” : their meaning refers to a directly observable or immediately “experienceable” factor: items abstracted from a wider, immediately apprehended or directly “inspectable” context. To abstract does not mean here to postulate, but to consider some feature apart. As opposed to concepts “by intuition” or “by inspection”, the meaning of a scientific concept is gained by virtue of the properties or relations assigned to it by postulates in a deductive theory of which the concept is an integral part (e.g., the concept of “blue”, which refers to a wavelength in electromagnetic theory, is a postulated concept that cannot be equated to the concept-by-intuition, “blue”). Consequently concepts by postulation essentially refer to theoretical entities, and not to entities that are part of phenomenal reality.

What interests us here is the fact that concepts by intuition or by inspection refer to factors that are operating in the phenomenal world. Thus their properly ‘observable’ nature is not in question. The belief of a social actor is, for example, an abstract factor that is not directly observable, but which can account for a decision, if the situation of the individual is such that he or she has no reason to put that in doubt in the furtherance of his or her goals. That which will for us characterize concepts by intuition or by inspection is the ‘experienceable’ character of the factors they refer to, i.e., the traceability of their effects. Northrop emphasizes that it is logically impossible to deduce from theoretical hypotheses any propositions that refer directly to observable entities. It is just as impossible, logically, to deduce from theoretical hypotheses any proposition whatsoever that refers directly to factors that are operating within reality. We are thus led to ask the following question: under what condition(s), from a formal model – elaborated in terms of postulated factors referring to entities that by construction are non-operating in the phenomenal world – may we derive true propositions regarding the production of a given phenomenon?

The response of Northrop is that connections between the formal postulated world and the phenomenal world can only be guaranteed by epistemic correlations. Epistemic correlations allow scientific concepts to have empirical meaning by linking the entities postulated via scientific concepts to concepts by intuition or by inspection that denote real entities or factors (Northrop 1947:143-144). Epistemic correlations should not be confused with the usual sort of correlation found in the sciences. The latter connects elements that belong to the same world. Epistemic correlations connect elements that belong to different worlds: scientific concepts (formal world) and concepts by inspection (phenomenal world). The fact of connecting two different modes of knowledge is precisely what qualifies these connections as epistemic. An important stage of scientific research should depend on the specification of relations, termed epistemic correlations, joining the concepts by postulation of the deduced theorems to the concepts by intuition or inspection which denote factors operating in the phenomenal world. Any claim on behalf of theoretical entities that they can reveal the truth about these factors is here set aside, even though the strong sense of explanation is being supported. Theoretical factors allow us to account for a particular role of their real correlates in a given process. The problem of discontinuity of the “reference” of scientific concepts from one theory to another appears to have no purpose here. The notion of epistemic correlation tells us that the relations between the unobservable, scientifically constructed, and the data of experience, do not send us back to relations of identity, nor to “correspondence rules”[11] as in the reductionist and unsuccessful version given by the logical positivists, but they are akin to relations of correlation. This interpretation of the relation between model and reality agrees with the representation of scientific theories as sets of models, but it is based on the existence of real correlates for at least some postulated theoretical factors, and thus takes seriously the question of explanation in a strong sense. It satisfies the exhortation by Hacking (1983 :419) that we should interest ourselves in this regard, translating this into the terms borrowed from

Northrop, not in the “truth” about the theoretical entities but in the traceable effects of their real epistemic correlates.

A formal model, representing theory Y for purposes of explanation of phenomenon X, can claim to provide support for true propositions regarding the production of X, and not merely an interpretative description of it, if epistemic correlations may be assumed. The truth of propositions under consideration depends on the validity of these correlations, which connect theoretical factors designated by scientific concepts to concepts by inspection which denote factors operating in the phenomenal world.

At this point, neither the clarifications offered by Northrop, nor those of Margenau allow us to account in a satisfactory manner for the problem of modelling of social action knowing that, in order to explain a social phenomenon, it is impossible to isolate in formal terms explanatory factors from other factors involved. We now turn to this problem in order to establish the conditions upon which the validity of epistemic correlations is supposed to rest.

4. The Modelling of Complex Phenomena Involves Hypotheses That Are Essentially Descriptive

From the standpoint of accounting for the role played by explanatory factors in the production of a social phenomenon X, over and above hypotheses related to the explanatory core of a theoretical model, complementary or secondary hypotheses, mainly purely descriptive ones, are put forward. The latter are capable of leading to false propositions regarding aspects of a phenomenon that are not the main focus of the model (Mäki 2000:328; Bailer-Jones, 2003:69). Such complementary hypotheses may be “construct idealizations,” such as Ernan McMullin describes them (McMullin 1985). In physics, some well-known examples of construct idealizations are frictionless surfaces, simplification of shapes, false vacuum hypothesis, etc. For the social sciences, the term “idealization”, understood as “a literally false exaggeration that serves an abstractive or isolating theoretical purpose” (Hausman 1992 :132) does not allow us to grasp all the forms of simplification that are involved. Agents that are motivated only by their own interests, perfect information, perfect markets, the homogeneity of goods, are for instance, in economics, well-known examples of idealization. In our view, nonetheless, they may not constitute “exaggerations” but in more general terms purely descriptive hypotheses, used in order to make sense of the effects of a complex grouping of factors.

Complementary or purely descriptive hypotheses refer not only to construct idealizations, but also to parts of the model whose aim is to represent phenomena, but not to explain the aspects of reality they grasp. In the presence of tightly interwoven mechanisms, you do not try to reproduce secondary factors and processes as they effectively operate, even in a simplified manner. In place of the numerous factors that have effects in reality, purely descriptive hypotheses aim only to represent main effects, without reference to epistemic correlation between ideal factors involved and real ones.

We may add that purely descriptive and explanatory hypotheses are necessarily interwoven at different levels and that, as stated above, explanatory status is relative. An example given by Margenau (1950:168-169) may illustrate this point. This example concerns Mendel’s laws of genetics which were deductively fertile but of a purely descriptive nature. In contrast to Mendel’s descriptive laws, Margenau explains, the modern theory which locates the genes within material carriers (the chromosomes) is looked upon as an explanation. It answers the question why hereditary traits are transmitted in certain ways, whereas Mendel’s laws merely show how this occurs. As Margenau puts it, the “why” is nothing more than a disguised “how”, in the sense, we may add, that the real factor involved, the chromosome, is not reducible to its epistemic correlate in biological theory but the latter provides support for true propositions about laws of genetics. Nevertheless for this very reason, a logical hierarchy distinguishes the two theories; the "why" is logically prior to the "how": Mendel's laws can be deduced from the theory of genes.

In all the models that attempt to recreate the composition of large-scale actions out of individual ones, the decision-making processes themselves are seldom an object of the theory represented by the model. Potential formal correlates of factors underlying these processes are not made explicit. Their whole results are the only ones that count for explaining the dynamic of the phenomenon under study. Explanatory hypotheses assume certain fundamental relationships that real decision-making procedures allow us to establish between factors held to be exogenous, and factors modified by actors' decisions.

Let's take a simple example, Schelling's famous model which accounts for generative mechanisms of spatial segregation. The basic model uses the movement of pieces on an 8 x 8 checkerboard and two kinds of pieces representing the members of two specific groups- men and women, blacks and whites etc. Pieces are distributed on the board leaving some empty squares. Then a decision process is defined in order to determine whether a piece will move. Such a decision about moving is supposed to depend on its discontent regarding its proximate neighborhood- i.e. pieces situated on the adjacent squares around. We suppose for instance that one piece leaves its location for the nearest satisfactory vacant square when more than two-thirds of the pieces around it are of a different kind. This hypothesis is assumed to sum up the operation of a set of factors (ethnic or social composition of a neighborhood, but also the kind of neighborhood, cost of housing, etc.) that may explain decisions to move. The decision algorithm in the model does not distinguish different motives. It uses a purely descriptive hypothesis that attributes the movement of a piece to the percentage of pieces of a different color in its vicinity. Other purely "descriptive" elements of the model can be cited which are of secondary importance: intensity of basic discriminatory motives, distribution of groups in space, definitions of neighborhoods, rules regarding individuals moving around, etc. The basic explanatory hypothesis is based on the structure of interdependence of decisions, in which individual actors (the pieces on the board) are decision-makers for themselves and components of the environment for their neighbors. The choice of an environment involves a process of endogenous change regarding the distribution of groups in space, where each decision to move modifies the environment of one's former neighbors as well as the new ones. This hypothesis is capable of participating in the explanation of all the social phenomena this same structure of interdependence applies to.

When the purpose of a model is to represent underlying factors that generate a phenomenon, this aim rests upon the primary hypotheses. Secondary hypotheses can be (and often are) essentially descriptive, for reasons having to do with simplification and concentration on the explanatory core of a model.

5. The Epistemic Criteria of Explanation

We are now prepared to respond to the central question here: under what conditions can formal models of social action claim explanatory power? We previously stated:

(1) A scientific explanation Y of a phenomenon X is intended to adequately describe the phenomenon X in terms of underlying factors that bring it about ;

(2) The validity of scientific explanation Y represented by a formal model depends on the possibility of deriving true propositions from the model regarding the production of phenomenon X, which is represented ;

(3) The truth of the propositions involved depends on the validity of the epistemic correlations assumed to connect formal factors to real factors assumed to produce X.

This presentation is of general application, as soon as we admit that explanation in the strong sense is distinguished from mere interpretation by the existence of real epistemic correlates for formal explanatory factors. This existence expresses the capacity of the model to give an account of the role played by generative factors, not its mere capacity to reproduce the phenomenon under study. The weight of explanation rests upon the accuracy of the role attributed to explanatory factors. Moreover the conditions of such accuracy cannot be translated by precise empirical consequences unless the interplay of explanatory factors can be observed in an isolated situation, or can be isolated in theory when the other factors involved interfere mechanically.[12] Thus we find ourselves nearer to the idea of a "nature of the process" [13] in Cartwright than to the conditions expressed by McMullin (1985:257) according to which the theory and its associated models "explain" the phenomenon in question (or the target system) if the explanatory hypotheses of the model ('theoretical laws' whose warrant is the explanatory theory) allow us to simulate the empirical regularities whose explanation is sought ('empirical laws' whose immediate warrant is observation or experiment).

As a result the validity of explanation Y depends on two major conditions being satisfied: the (indirect) reference of theory Y to factors that actually operate in the phenomenal world, and the relative accuracy of the effects that are attributed to those factors within the developed formal framework. What we are maintaining is that these two conditions entail two more which are commented on just below. In the first place, the modelling of underlying processes that generate phenomenon X presupposes the "abstraction" of factors that produce X, and their formal representation, but not their ideal isolation (Lawson 1997: 131-133; Long 2004). In the second place, hypotheses that are purely descriptive and the explanatory hypotheses of the model must be relatively independent with regard to the represented phenomenon, X.

(1) The modelling of underlying processes that generate phenomenon X *under study translates the effects of real factors abstracted from their phenomenal context, not those of purely ideal ones.*

A model that is intended to provide an explanation must transcribe the set of mechanisms that genuinely underlie the production of the social phenomenon X. On this view, the explanatory formal factors are assumed to be connected to factors operating in the phenomenal world by means of epistemic correlations. The correlates of formal explanatory factors, operating in the phenomenal world, are abstract in the sense that they represent common traits of the category of phenomena to which their action is imputed. This abstraction should be the highest possible: they denote the common traits of all the factors that are susceptible of producing the effect attributed to them. This level of abstraction is achieved when the abstract factor identified cannot be replaced by another that in reality shares with it a common trait (more abstract) which is the true real factor underlying the imputed effect.

Defined via an operation of abstraction, factors involved are not ideal factors. For example, agents that are omniscient, infallible, and completely free to act, and economic actors that maximize their own utility, are non-realistic hypotheses. If they refer to the action of ideal factors, they are not part of the explanatory hypotheses whose function is to account for the effects of factors operating in the phenomenal world. As stated by Long (2004), it is one thing to omit specifying details of reality because they do not account for the phenomenon under examination, and something else again to specify their absence: absence of error in the case of the optimizing individual, absence of limits on the circulation of information, etc. From an explanatory point of view, it is necessary to distinguish clearly between abstraction in the sense of considering factors apart from the larger context in which they operate, and the creation of fictitious factors.

Note that certain aspects of the real cannot be ignored when we are considering the individual action of factors involved, but they can be when the results of joint action are being considered. This observation greatly increases the possibility of epistemic correlations between real and formal factors. In particular, the hypothesis of rationality does not presuppose that all social actors act in a rational manner all the time, but only that their tendency to act rationally is the most important common factor (Goldthorpe 2000 :116; Mäki 2000 :323).

This fundamental tendency or capacity is not thought of here in a narrow utilitarian sense, but in a larger cognitive sense, putting in play a set of complex processes that are potentially transcribed in terms of situations (positions within a system of action, and cognitive dispositions in particular) in which social actors find themselves. The principle of rationality plays in this respect a role that is primarily one of arbitration. It allows us to judge the likelihood of decisional processes whose results underlie social actors' behavior, and then to evaluate the relative pertinence of competing alternatives. The « law of large numbers » insures that, for each specific situation identified, the common tendencies predominate. Formal factors, which only embody the dominant tendencies common to individual actions, are intended to transcribe effects of factors operating conjointly in the phenomenal world. Even if the behavior of profit-maximizing businessmen does not correspond to the reality of the experience of these economic actors, they still hypothetically represent a leading tendency in response to competitive markets. Particular situations can be ignored to the extent that they have no systematic effect on the tendencies that are being represented. The specific factors are not in such a case considered as absent, since if they were introduced, their effects would cancel each other out mathematically. The variability of decision-making contexts can also account for the emergence of that which Miller and Page (2007: 49-52) understand by "organized complexity", in which interactions between actors, rather than canceling one another out, reinforce one another. In this regard, the tendency of agent-based models: ABMs to seek “ laws” of behavior that would permit one to simulate emergent phenomena at the social level, should not deceive us concerning the meaning of these “ laws”. The laws in question have descriptive qualities that provide support for the explanation of the dynamics of the phenomenon under study. They do not reveal behavioral tendencies as such. They might allow us to account for the phenomenon of emergence, but on the other hand they have no explanatory value regarding the behavior of the agents involved. The effects described as social influence or imitation could for example be explained by latent factors involving consequences resulting from changes in situational parameters of individual decisions that have nothing to do with mere influence or imitation.

In opposition to this claim, there is a reductive hypothesis common to ABMs that stipulates that agents usually follow simple rules in the sense that they are not supposed to be cognitively complex (Macy & Willer, 2002:146). The reason for such a hypothesis may be that it offers a theoretical base, compatible with an evolutionary viewpoint, for experimental explorations without accessing a base of empirical data. These explorations concern, for example, the problem of the emergence of norms or values. From this perspective, agents are characterized as “adaptive” (Holland, 1995), which supposes that the actions in view involve mostly imitation and learning, learning being attributed to evolutionary processes of selection, imitation, and social influence.

Let’s take the example of the development of cooperation such as this is explained by the politologist Robert Axelrod (1984) on the basis of a confrontation of strategies in a iterative Prisoner Dilemma game. The simplest of all rules, tit-for-tat, was the winning strategy in a tournament where each strategy proposed by a wide range of social scientists was matched against every other one (Robert Axelrod, 1984). We understand by strategy the specification of what to do in any situation that might arise (Axelrod 1984 :14). Following tit-for-tat, the actor begins by cooperating with its opponent. It then plays exactly as the latter had played in just the previous game. If the latter had defected, the actor also defects. If the opponent cooperates, the actor cooperates. Despite the simplicity of this strategy, its adoption as a systematic rule is based on a sharp analysis of the game, which we suppose to be played over an infinite number of times. It was proposed for the tournament by a specialist, the mathematical psychologist Anatol Rapoport. The use of these results in the explanation of the role played by cooperation in an evolutionary dynamics[14] raises a fundamental issue: the reality of decisions does not respond to systematic strategies over the long term, but does respond to successive evaluations of particular situations. In this respect if tit-for-tat is not realistic, from an explanatory point of view, as a long term strategy, it is not realistic as a short term systematic rule for action, either. As Axelrod (1984:39-40) explains, a strategy that would easily have won the first part of the contest if it had been tried is a variant of "Downing's one" (1975) which views Prisoner Dilemma behavior as problem-solving behavior. The rule is based on an attempt by the player to understand the responses of the other player to his or her choices and to make an adequate decision based on this understanding.[15] Note that this strategy was originally developed as an interpretation of the choice of subjects in Prisoner's Dilemma laboratory experiments. The second part of the contest revealed that the variants of Downing proposed were dominated by specific programs expressly defined in order to take advantage of the others. Alternative hypotheses attached to explanatory realism of factors represented lead simply to the suggestion that evolution produced a force for cooperation with evaluative and strategic potentialities far superior to that produced by the tit-for-tat rule: *rationality* (in the cognitive sense).

Now let us turn to the second condition of explanation, a condition applied to the validity of epistemic correlations supposed to exist between formal explanatory factors and real ones. As stated above, the adequacy of results regarding a given phenomenon cannot assure the adequacy of the role imputed to factors involved via their supposed formal correlates within the model in cases where they cannot be isolated in reality. The meaning of formal factors and their effects are a matter of their relationship with other factors that are accounted for in the model, not with the reality that is represented. The formal representation of the effects of main factors imposes the presence of hypotheses that only aim to give form to the representation of phenomena under study – hypotheses that are essentially descriptive. Whence we draw this formula for the second condition of the validity of the claim of a model to explanatory power:

(2) The validity of an explanation based upon or supported by a model depends on the relative autonomy of explanatory hypotheses Y and purely descriptive hypotheses in relation to the reality represented, X.

The condition of relative autonomy between strictly descriptive hypotheses and explanatory hypotheses does not only aim at the « stability » of a model, that is, its non-sensitivity to artifactual effects of secondary hypothesis. This condition applies more generally to the validity of supposed epistemic correlations between formal factors and real factors (through their observed effects), in other words the validity of the explanation itself. The meaning that we assign to this condition is defined as follows. Hypotheses that are essentially descriptive (whose function is to operate conceptual simplifications and to account for the effects of complex processes that are not made explicit by the model) and explanatory hypotheses Y (which account for the effects of factors that generate X) are relatively autonomous with regard to phenomenon X, if no latent factor interferes with supposed real correlates of explanatory formal factors, to a degree that would invalidate the explanation offered, Y.

In other words, it must theoretically be possible to develop purely descriptive assumptions in order to render more explicit the underlying process they describe, introducing previously ignored factors and opening up the black boxes, all without falsifying the effects previously attributed to the explanatory factors of the model in the production of X.

At any rate, the postulated independence between the two types of hypotheses, i.e. the validity of Y, depends on the level of generality at which the model operates. That is why the independence referred to is only relative. Latent connections must exist between explanatory hypotheses and descriptive ones, taking into account the complexity of the phenomena modeled, but these connections can be ignored if they do not falsify the conclusions derived from the model. Conversely if the condition of relative autonomy is not fulfilled, the role attributed to explanatory factors within the model is skewed.

We must add that this condition is at least an implicit hypothesis of the theory Y. But respect for it supposes many more justifications than are usually given. It fulfills a requirement of theoretical research, which the confrontation of emergent properties of formal systems with empirical data can never guarantee. The validity of epistemic correlations assumed to exist between real and formal factors is a matter of the analytic relevance of the generative mechanisms represented. It cannot theoretically be tested, except by comparison with all conceivable alternate theories; such a test being consequently never definitive.

In practice, to test the robustness of explanatory hypotheses of a model, one method is to experiment with developments of the purely descriptive hypotheses, aiming at invalidating the explanatory core of the model. This leads to « attacking » the generative mechanisms whose behavior is modeled by pulling the explanation in the direction of other factors. It is a question of putting forward alternatives to the model to test the validity of epistemic correlations that are postulated to exist between explanatory factors in the model and real factors involved. This test leads to a comparison of potential interpretations of real processes, and to the evaluation of the pertinence or applicability of the various simplifications the model involves. The explanatory core constitutes the target for these alternative interpretations of generative mechanisms. It expresses the level of generality assigned to the explanation Y in relation to the represented phenomenon, X, and it is by situating our examination at that same level that we are able to analyze the robustness of Y.

This effort to test the explanatory power of a model can to some degree be likened to the procedure of de-idealization set forth by McMullin (1985 :259) for the validation of a developed theory. Simplifying and other non-realist hypotheses are "relaxed" by gradually reintroduce the complexities within the model. This requires, writes McMullin, a knowledge of how that particular 'complexity' operates. This operation has to be guided by the explanatory theory because the idealization is supposed to have been so guided. But the approach taken here is more general. It is not empirical adequacy that plays the role of validation criterion of the explanatory potential of the model, but the adequacy of epistemic correlations, i.e., the validity of effects imputed to the factors involved via their correlates within the model. The possibility of experimentally isolating the factors operating allows us to insure an empirical adequacy that will eventually be strong; but that is never the case in the social sciences. In order to confirm the explanatory power of a model, the fundamental procedure is not to verify correspondence with observed data by relaxing the complexities, but rather to augment the explanatory or analytical realism of certain simplifying hypotheses in order to test the structure of formal explanation.[16]

As we have already underlined, in social action models that aim at accounting for the effects of the composition of individual actions, decision making processes are modeled in a general descriptive manner. The decision procedure is extremely simplified in order, essentially, to account for the effects of explanatory factors on the overall results of the aggregation of decisions. But only a comparison with alternatives, based on developments in decision-making processes from the standpoint of analytical realism, can allow us to properly judge the analytic relevance of the model.

Let us return to the example of Schelling's model of urban segregation. The model shows that individual motivations laying behind a phenomenon of segregation might be much weaker than the degree of segregation observed. The explanatory core of the model relies on a process of amplification based on the structure of interdependence previously outlined: people are decision-makers for themselves and components of the environment for their neighbors. Schelling observes that it is not easy to determine the boundaries between segregation due to « individual » motivations, and other types, such as segregation due to economic conditions. Even the latter is linked to discrimination to the extent that "choosing a neighborhood means choosing your neighbors" (Schelling 1971:139). But in reality the explanation of the phenomenon of residential segregation in this alternative hypothesis is quite different: it does not rely on criteria of appreciation directly involving neighbors nor interdependence of decisions. The factors that explain individual preferences might turn out to be transcultural and be affected primarily by situational constraints. If the satisfaction function is modeled using a Cobb-Douglas function[17] from this point of view, it does not reveal a split between individual preferences and collective results, but quite to the contrary, there is progressive convergence between the two. Spending 20% of your income on housing amounts to choosing a neighborhood where a certain level of income is common. Therefore we are no longer in an interpretive context that centers on a sociocultural conflict. The various mechanisms leading to forms of segregation may merge in order to account for given social phenomena. It is in regard to particular types of contexts that the effects imputed to generative mechanisms can be tested. In other words, the explanatory power of a model can only be evaluated in reference to the underlying regularity subsuming a set of specific social realities. This may lead to distinguish contexts in regard to the generative mechanisms involved. The realist development of the processes of individual decision making suggests us here that explanatory assumptions inspired by Schelling's model should be to a greater or lesser extent completed in numerous cases of urban segregation with assumptions relative to economic motives. The basic problem raised here does not concern the trivial question of the addition of a factor that will eventually be neglected, but rather the falsification of the explanatory structure of a model through a realistic development of hypotheses that are essentially descriptive (here concerning decision processes). We should add that, as the meaning of formal factors depends on their relations with other formal factors, tests of alternatives through the development of implicit factors allow us to better interpret their role with regard to the reality that is represented.

Note that satisfying the condition of relative independence of descriptive and explanatory hypotheses, i.e. the validity of effects imputed to explanatory factors in the generation of the phenomenon, could prove difficult when models allow a large number of degrees of freedom, which is a potential problem with agent-based models - knowing for instance that the initial situation of each agent in the model may influence the results of the simulation (Hegselmann & alii. 1996; Miller & Page 2007; Amblard & Phan 2007). One solution offered by Miller and Page (2007:75) has to be followed: running the model a hundred or a thousand times in order to obtain the distribution of possible outcomes.

6. Conclusion

Finally, models only explain in so far as they deploy the underlying factors bringing about the phenomena represented. It is a matter of explanatory or analytical realism. The importance of permitted simplifications rests upon the generality of the phenomena explained - i.e. of the possible range of concrete social systems sharing the same explanatory structure. These simplifications complete a movement of abstraction which, as specified above, retain only those factors that genuinely produce the phenomena being analyzed, and they sum up complex secondary mechanisms by means of descriptive hypotheses. When realist developments (in the explanatory sense) of descriptive hypotheses do not lead to invalidating the explanatory core of a model, they prove to be useless. This test assures us that a level of simplicity, that is as great as possible, does not affect the validity of the explanation. Nonetheless, it is possible to hold that such a test is never completely definitive, to the extent that the generative underlying factors rely on representation of reality that is always open to scientific analysis.

Acknowledgements The author would especially like to thank the editor and two anonymous referees of ISPS for their very constructive comments and suggestions on earlier version of this paper.

REFERENCES

Amblard, F., Phan, D. (2007) Agent-based modelling and simulation in the social and human sciences (London, Bardwell Press).

Axelrod, R. (1984) The evolution of cooperation (Cambridge, Basic Books).

Bailer-Jones, D.M. (2003) When scientific models represent, International Studies in the Philosophy of Science, 17, pp.59-74.

Bunge, M. (1997), Mechanisms and Explanation, Philosophy of the Social Sciences, 27, pp.410-465.

Boudon, R. (1998), Social mechanisms without black boxes, in: P.Hedström and R.Swedberg (Ed.) Social mechanisms. An analytical approach to social theory (Cambridge, Cambridge University Press).

Boudon, R. (2003) Beyond rational choice theory. Annual Review of Sociology, 29:1-21: 172-203.

Cartwright, N. (1983) How the laws of physics lie (Oxford, Clarendon Press).

Cartwright, N. (1989) Nature’s capacities and their measurement (Oxford, Clarendon Press).

Cartwright, (1999) The dappled world. A study of the boundaries of science (Cambridge, Cambridge University Press).

Cherkaoui, M. (2005) Invisible codes: essays on generative mechanisms (Oxford, The Bardwell Press).

Downing, L.L. (1975) The Prisoner’s dilemma game as a problem-solving phenomenon: an outcome maximizing interpretation, Simulation and games, 6; pp.366-91.

Ellis, B. (1996) What science aims to do, in: D. Papineau D (Ed.) The philosophy of science (Oxford, Oxford University Press): 166-193.

Elster, J. (1998) A plea for mechanisms. in: P.Hedström and R.Swedberg (Ed.) Social mechanisms. An analytical approach to social theory (Cambridge, Cambridge University Press): 45-73.

Epstein, J.M., R.Axtell (1996) Growing artificial societies: social science from the bottom up (Washington D.C. Brookings Institution Press).

Epstein, J.M. (2007) Generative social science. Studies in agent-based computational modelling (Princeton, Princeton University Press)

Fraassen, van B.C. (1980) The scientific image (Oxford, Clarendon Press).

Giere, R.N. (1988) Explaining science. A cognitive approach (Chicago, The University of Chicago Press).

Goldthorpe, J.H. (2000) On sociology. Numbers, narratives and the integration of research and theory (Oxford University Press).

Hacking, I. (1983) Representing and intervening: introductory topics in the philosophy of natural science (Cambridge, Cambridge University Press).

Harré, R. (1960) Metaphor, model and mechanism, Aristotelician Society, pp.101-122.

Hausman, D. (1992) The inexact and separate science of economics (Cambridge, Cambridge University Press).

Hedström, P. and Swedberg R. (1998) Social mechanisms. An analytical approach to social theory (Cambridge, Cambridge University Press).

Hedström, P. (2005) Dissecting the social. On the principles of analytical sociology (Cambridge, Cambridge University Press).

Hegselmann, R., Mueller U., Troitzsch K.G. (Eds.) (1996) Modelling and simulation in the social sciences from the philosophy of science point of view (Dordrecht/ Boston / London, Kluwer Academic Publishers, Theory and Decision Library).

Holland, J.H. (1995) Hidden order. How adaptation builds complexity (New York, Basic Books).

Lawson, T. (1997) Economics & reality (London, Routledge).

Levins, R. (1966) The Strategy of model Building in population biology, American Scientist, 54, pp.421-431.

Lévy-Lambert, H. (1969) Modèle de choix en matière de politique du logement, Revue d'économie politique, LXXVIII, pp.938-965.

Long, R. (2004) Realism and abstraction in economics : Aristotle and Mises versus Friedman, Austrian Scholar Conference, 10.

Macy, W.M., Willer R. (2002) From factors to actors: computational sociology and agent-based modelling. Annual Review of Sociology, 28:143-166.

Mäki, U. 2000. Kinds of Assumptions and their truth: Shaking an Untwisted F-Twist. *Kyklos*, *53*: 317-336.

Mäki, U. (2003) 'The methodology of positive economics' (1953) does not give us *the* methodology of positive economics, Journal of Economic Methodology, 10, pp.495-515, 2003.

Manicas, P.T. (2006) A realist philosophy of science. Explanation and understanding (Cambridge, Cambridge University Press).

Margenau, H. (1935) Methodology of modern physics, Philosophy of Science, 2, pp.48-72 et 164-187.

Margenau, H. (1950) The nature of physical reality. A philosophy of modern physics (New York, McGraw-Hill Book Company)

Mayntz, R. (2004) Mechanisms in the analysis of social macro-phenomena, Philosophy of the Social Sciences, 34, pp.237-259.

McMullin, E. (1985) Galilean idealization, Studies in History and Philosophy of Sciences, 16, pp.247-273.

Mill, J.S. (1836) A system of logic (New York, Harper & Brothers).

Miller, R. (1987) Fact and method. Explanation, confirmation and reality in the natural and the social sciences (Oxford, Princeton University Press).

Miller, J.H., Page S.E. (2007) Complex adaptative systems. An introduction to computational models of social life (Princeton, Princeton University Press)

Da Costa, N.C.A., French, S. (2003) Science and partial truth. A unitary approach to models and scientific reasoning (Oxford, Oxford University Press).

Northrop, F.S.C. (1947) The logic of the sciences and the humanities (The Macmillan Company, New York).

Schelling, T.C. (1971) Dynamic models of segregation, Journal of Mathematical Sociology, 1, pp.143-186.

Suarez, M. (2008) Experimental realism reconsidered: how inference to the most likely cause might be sound in: L. Bovens and S. Hartmann (Eds.) Nancy Cartwright's philosophy of science (London, Routledge) pp.137-163.

Sugden, R. (2000) Credible worlds: the status of theoretical models in economics, Journal of Economic Methodology, 7, pp.1-31.

Stinchcombe, A.L. (1991) The conditions of fruitfulness of theorizing about mechanisms in social science, Philosophy of the Social Sciences, 21, pp.367-388.

Weisberg, M. (2006) Forty years of 'The Strategy': Levins on model building and idealization, Biology & Philosophy, 21, pp.623-645.

NOTES

[1] See Stinchcombe (1991); Bunge (1997); Hedström and Swedberg (1998), and especially Boudon (1998) and Elster (1998) therein; also Goldthorpe (2000), Mayntz (2003), Cherkaoui (2005), Hedström (2005) and Manicas (2006).

[2] For instance, according to Sugden (2000), the model of inductive inference for explanation is:
(E1) In the model world, *R* is caused by *F*
(E2) *F* operates in the real world
(E3) *R* occurs in the real world
Therefore, there is reason to believe:
(E4) In the real world, *R* is caused by *F*.

[3] While in numerous models used to simulate the outcomes of the composition of multiple social actions, the interdependence between decisions is indirect - i.e. affecting situations and not decision rules as such, in agent-based models, the effects of interaction between actors tend to affect decision-making processes themselves. These ones are then involved in a dynamic process of change, such as occurs with the phenomena of learning or influence. Whereas we speak of agent-based models as soon as interdependence of actions is involved, the latter approach tends to characterize more recent trends of research.

[4] See, for example Margenau (1935, 1950); Harré (1960); Achinstein (1968); Cartwright (1983; 1989); McMullin (1985); Miller (1987); Ellis (1996); Bunge (1997); Mäki (2003); Bailer-Jones (2003); Mayntz (2003); Cherkaoui (2005).

[5] Note that causal claims were rejected from the beginning by the supporters of positivism.

[6] This version of the nature of explanation is presented, for example, by Newton da Costa and Steven French (2003:59). Explanation is supposed to rest on complex series of connections between models that are interwoven and organized into a hierarchy, running from "data" models and "phenomenological" models to their "high-level" theoretical counterparts. These connections are represented in terms of partial isomorphisms "holding between families of partial relations in the structures constituting these models".

[7] Ronald Giere also adopts a semantic interpretation of theories according to which (1988:85) a theory comprises two elements: (1) a population of models, and (2) various hypotheses linking those models with systems in the real world. According to the "constructive realism" that he defends (1988:97) "theoretical hypotheses are interpreted as asserting a similarity between a real system and some, but not necessarily all, aspects of a model."

[8] Mill notes that it is not quite correct to say that a body moves in a certain way because of the action of a certain force; unless it is prevented from doing so by another force, it tends to move in the same way even when it undergoes the influence of another force (cf. Cartwright 1989: 177).

[9] The "DBO" theory sustained by Hedström (2005) is founded on the assumption that Desires (D), Beliefs (B) and Opportunities (O) are the primary theoretical terms upon which the analysis of action and interaction could be based. This theoretical framework is in our view of less scope than the concept of cognitive rationality (which more generally supposes motives, cognitive capacities, notably reflexive ones, and situations, all accounting for the actor's reasons for acting). Making no assumption that actors act rationally but assuming they act reasonably and with intention (Hedström 2005: 61), DBO theory overshadows the human meaning of cognitive rationality and imposes a reductive theoretical framework.

[10] Understood in a less specific sense than the notion of tendency in Mill (1836) or capacity in Cartwright where the notion plays the role of the *explanans* and does not send us back to actual results when different factors interact. The tendency here represents the *explanandum*, expressed in a theoretical or observational language that makes use of units of analysis generally situated at a higher level than units of the *explanans* (Stinchcombe 1991).

[11] Note that Margenau speaks of "correspondence rules" instead of "epistemic correlations". But as he observes (1950:57): "logical positivism, in so far as it restricts itself to an analysis of scientific language, can never do complete justice to science ; it must forever talk about propositions, whereas the scientist concerns himself with meanings that are prior to propositions".

[12] Using Mill's terminology, when causal factors combine "mechanically", the effect of their combination is the same as the total effect of each acting separately. See for instance Hausman (2003:137) and Lawson (1997:132).

[13] For example, if we ask what can be generalized from a specific experimental result, "The answer requires the notion of nature: the features that are necessary are exactly those which, in this very specific concrete situation, allow the nature of the process under study to express itself in some readable way." (Cartwright 1999:90).

[14] In reality, tit-for-tat never totalises more, and over the long term never less, than the adverse strategy. Nonetheless, any defection by the adversary brings about a reaction of defection by the player that cannot be overcome, except through the cooperation of the adversary.

[15] The winning version is based on a starting hypothesis that is optimistic rather than pessimistic, with regard to the reactivity of the other player. It is rational in the case of an iterative game in this sense, that even if a player does not know a priori the profile of the adversary, he knows that interaction with him will be repeated a potentially infinite number of times.

[16] The procedure is closer to that proposed by Richard Levins (1966), which suggests developing several alternative models to test the core causal structure of a model against the development of the details left out by simplifying assumptions. Nonetheless, what is being discussed here is of greater generality. Hypotheses that are essentially descriptive do not only bring about causal simplifications through "magnification", which correspond to those that Levins had in view (see also Weisberg 2006).

[17] For example $S(q,x) = q^{(1-a)} x^{a}$ with $pq + lx = r$ where " x" stands for "consumption of housing" translating the quality of the housing ; " q" stands for the importance of other items consumed in the household ; " a" is a measurement of the preference for housing ; " lx" stands for the annual charge corresponding to housing ; " p" stands for the price index for other uses for income ; " r" stands for the household's income. Satisfaction is supposed to be maximal when the portions of income of a household that are spent on housing and on other consumptions are equal to "a" and "1-a" (Levy-Lambert 1969).