1

Introduction

In the 1950s, all hadrons, namely particles that are involved in strong interactions, including the proton and neutron (or nucleons) and other baryons, together with pions and kaons and other mesons, were regarded as elementary particles. Attempts were made to take some particles, such as the proton, neutron and lambda particle, as more fundamental than others, so that all other hadrons could be derived from the fundamental ones (Fermi and Yang, 1949; Sakata, 1956). But the prevailing understanding was that all elementary particles were equally elementary, none was more fundamental than others. This general consensus was summarized in the notion of "nuclear democracy" or "hadronic egalitarianism" (Chew and Frautschi, 1961a, b; Gell-Mann, 1987).

As to the dynamics that governs hadrons' behavior in the processes of strong interactions, early attempts to model on the successful theory of quantum electrodynamics (or QED, a special version of quantum field theory, or QFT, in the case of electromagnetism), namely the meson theory, failed, and failed without redemption (cf. Cao, 1997, Section 8.2). More general oppositions to the use of QFT for understanding strong interactions were raised by Landau and his collaborators, on the basis of serious dynamical considerations (Landau, Abrikosov, and Khalatnikov, 1954a, b, c, d; Landau, 1955). The resulting situation since the mid 1950s was characterized by a general retreat from fundamental investigations to phenomenological ones in hadron physics. The prevailing enquiry was phenomenological because no detailed understanding of what is going on in strong interactions was assumed or even aspired to, although some general principles (such as those of crossing, analyticity, unitarity, and symmetry) abstracted from some model dynamical theories were appealed to for reasoning from inputs to outputs; thereby the enquiry enjoyed some explanatory and predictive power.

At the end of the 1970s, however, none of the hadrons was regarded as elementary any more. The unanimous consensus in the physics community and, through popularization, in the general public became as follows. First, all hadrons were composed of quarks that were held together by gluons; and second, the dynamics of quark–gluon interactions was properly understood and mathematically formulated in quantum chromodynamics (or QCD). As to the strong interaction among hadrons, it could be understood as the uncancelled residual of the quark–gluon super-strong interaction, a kind of Van der Waals force of the hadrons.

Such a radical change in our conception of the fundamental ontology of the physical world and its dynamics was one of the greatest achievements in the history of science. The intellectual journey through which the conception was remolded is much richer and more complicated than a purely conceptual one in which some ideas were replaced by others. The journey was fascinating and full of implications, and thus deserves comprehensive historical investigation. However, even the conceptual part of the story is illuminative enough to make some historical and philosophical points.

While a full-scale historical treatment of the episode is in preparation, (Cao, forthcoming) the present enquiry, as part of the more comprehensive project, has a more modest goal to achieve. That is, it aims to give a concise outline of crucial conceptual developments in the making of QCD. More precisely, its attention is restricted to the journey from the proposal of current algebra in 1962 to the conceptual and mathematical formulation of QCD in 1972–73.

As a brief conceptual history, its intention is twofold. For the general readers, it aims to help them grasp the major steps in the reconceptualization of the fundamental ontology of the physical world and its dynamics without being troubled by technical details. However, it is not intended to be a popular exposition. For experts who are familiar with the details (original texts and technical subtleties), it promises to offer a decent history, in which distorted records will be straightened, the historical meaning of each step in the development clarified, and significance properly judged, on the basis of present understanding of the relevant physics and its historical development, that is, helped by hindsight and present perspective.

The preliminary investigations pursued so far have already revealed something of deep interest, and thus provided a firm ground for making some claims about the objectivity and progress of scientific knowledge, the central topics in contemporary debate about the nature of scientific knowledge and its historical changes.

Pivotal to the debate is the status of unobservable theoretical entities such as quarks and gluons. Do they really exist in the physical world as objective entities, independently of human will, or exist merely as human constructions for their utility in organizing our experiences and predicting future events? If the former is the case, then a related question is whether we can have true knowledge of them, and how? Thus the notion of unobservable entity is central to metaphysics, epistemology and methodology of theoretical sciences.

In the debate there are, roughly speaking, two camps. One is the realist camp, and the other antirealist. Realists take the objective existence of unobservable entities for granted if these entities can consistently give us successful explanation and predictions. They may differ in how to know the entities, but all of them are optimistic in human ability to know them. As a corollary, historical changes of scientific knowledge, according to realists, are progressive in nature. That is, the change means the accumulation of true knowledge of the objective world, consisting of observable as well as unobservable entities structured in certain ways. The necessity of the unobservable entity comes from the hypothetic-deductive methodology, which, in turn, has its deep roots in human desire for explanation.

For antirealists, the status of the unobservable entity is dubious at best. Antirealists find no justification to take it as more than a fictitious device for convenience. They refute the realist argument for its objective existence, mainly the success it has brought in explanation and prediction, as being too naïve, and deploy their own more "sophisticated" arguments, one logical, the other historical, to remove the notion of unobservable entities from our basic understanding of theoretical sciences.

The logical argument is based on the notion of underdetermination. The underdetermination thesis suggested by Pierre Duhem (1906) and W. V. O. Quine (1951) claims that in general no theoretical terms, and unobservable entities in particular, can be uniquely determined by empirical data. That is, given a set of evidence, we can always construct more than one theory, each of them based on some unobservable entities as its basic ontology for explanation and predictions; while all of these theories are compatible with the evidence, the hypothetical entities assumed by these theories may have conflicting features, and thus cannot be all true to the reality.[1] Once the logical ground for inferring the reality of unobservable entities from evidence is removed, the existential status of unobservable entities can never be settled, that is, their status can only be taken as conventional rather than objective.

It has been noticed that the convincing power of the Duhem–Quine thesis rests entirely on taking unstructured empirical data (or more precisely, structured in its existent form) as the sole criterion for determining the acceptability of a hypothetical entity. Once this kind of data is deprived of such a privileged status, the simplistic view of scientific theory as consisting only of empirical, logico-mathematical, and conventional components is to be replaced by a more sophisticated one, in which a metaphysical component (e.g. one that is responsible for the intelligibility and plausibility of a conceptual framework, which is the result of, and also a foundation for, a particular way of structuring data) is also included and plays an important role in selecting acceptable unobservable entities. This would surely put scientific theories in a wider network of entrenched presuppositions of the times and in a pervasive cultural climate, and thus invites cultural and sociological studies of science to join force with history and philosophy of science in our effort to understand scientific enterprise. If this is the case, the Duhem–Quine thesis alone is not powerful enough to discredit the realist interpretation of unobservable entities.

In the last four decades, however, the antirealist has relied more heavily on its historical argument, which is based on the notion of scientific revolution, a notion that was made popular mainly by Thomas Kuhn. If the Duhem–Quine thesis accepts the existence of a multiplicity of conflicting theoretical ontologies, and thus nullifies the debate on which ontology should be taken as the true one, Kuhn rejects the reality of any theoretical ontology: if whatever ontology posited by a scientific theory, no matter how successful it was in explanation and prediction, is always replaced, through a scientific revolution, by another different and often incompatible one posited by a later theory, as the history of science seems to have shown us, and there is no coherent direction of ontological development in the history of science, how can we take any theoretical ontology as the true ontology of the world (Kuhn, 1970)? If there is no reason to believe that there will be an end of scientific revolution in the future, then, by induction, the privileged status of the unobservable entities discovered or constructed by our current successful theories has to be deprived (Putnam, 1978).

Thus the rejection of the reality of unobservable entities is reinforced by the claim of discontinuity in history of science, which takes the pessimistic induction argument just mentioned as its most combative form. A corollary is that, according to antirealists, no claim to progress could be made in terms of accumulation of true knowledge of the objective world. The true role of unobservable entities, in which our knowledge is encapsulated, is not to describe and explain what actually exists and happens in the world. Rather, they are constructed for our convenience in making successful predictions.

A difficult question for the antirealist is: why some constructions are successful and others are not? The realist argues that if the success of science is not a miracle, then the successful theory and its hypothetical, unobservable entities must have something to do with reality. If simply taking every unobservable entity in a successful theory as what actually exists in the world, for the reasons raised by the antirealist, is too naïve an attitude, then at least one can argue that the relational and structural aspects of a successful theory must be real in the sense that some similar aspects exist in the world. If this is the case, then the connection between theory and evidence can be recovered and the continuity of scientific development can be properly argued for. This is the so-called structural realist position, a more sophisticated approach to realism indeed.

Structural realism was first conceived by Henri Poincare (1902), and then deliberated by Bertrand Russell (1927), Ernst Cassirer (1936, 1944) and others. In recent decades, it has been intensely pursued by Grover Maxwell (1970a, b), John Worrall (1989), Elie Zahar (1996, 2001), Steven French (2003a, b), Tian Yu Cao (1997, 2003a, b, c), and others. In its current incarnation, structural realism takes different forms. Common to all these forms is a recognition that a structure posited or discovered by a successful theory, as a system of stable relations among a set of elements or a self-regulating whole under transformations, in contrast with unobservable entities[2] underlying the structure, is epistemically accessible, thus its reality can be checked with evidence (up to isomorphism, of course, due to its relational nature), and the objectivity of our knowledge about it is determinable.

Apparently, structural realism smacks of phenomenalism. But it can be otherwise. A crucial point here is that a structure, while describing a recognized pattern in phenomena, such as those patterns recorded and suggested by global symmetry schemes for hadron spectroscopy, may also point to deep reality, such as quarks and gluons, both in terms of a deep structure underlying the patterns in phenomena, such as the one suggested by the constituent quark model of hadrons, and also in terms of hidden structuring agents that hold components together to be a coherent whole, such as permanently confined color gauge bosons. The conceptual development that will be recounted in the following chapters will illuminate this crucial point in a convincing way.

A vexing question for structuralism in all areas, perhaps with the exception of certain branches in mathematics, is that a structure in a scientific theory has relevance to the real world only when it is interpreted, usually by specifying the nature and properties of its underlying elements. Since a structure can be interpreted in different ways, we are facing underdetermination again. In addressing this vexing question, three different positions have emerged from structural realism.

The first position, known as epistemic structural realism, takes an agnostic attitude toward underlying unobservable entity, and restricts reliable scientific knowledge only to structural aspects of reality, which is usually encapsulated in mathematical structures, without involving the nature and content of underlying entities whose relations define the structure (Worrall, 1989). With such a realistic understanding of structural knowledge, this position has somewhat addressed the pessimistic induction argument: the history of science is nothing less than a process in which true structural knowledge is accumulated, and thus is continuous and progressive in nature. However, as far as unobservable entity is concerned, this position is not different from the antirealist one. For this reason, Kuhn's original claim, that there is no coherent direction of ontological development in the history of science, is evaded rather than properly addressed, if by ontology we mean the fundamental entities in a domain of scientific investigations, from which all other entities and phenomena in the domain can be deduced.

The second position, known as ontic structural realism, is extremely radical in fundamental metaphysics and semantics (French and Ladyman, 2003a, b). It claims that only structures are real, no objects actually exist; and that the phenomenological existence of objects and their properties has to be reconceptualized purely in structural terms. For example, electric charge has to be understood as self-subsistent and a permanent relation, and elementary particles have to be understood in terms of group structures and representations. By taking structures as the only ontology in the world, Kuhn's ontological discontinuity claim is addressed, and the continuity and progress in the historical development of science can be defended. But the price for these gains is that the very notion of unobservable entity is dissolved and eliminated altogether from scientific discourse.

The third version, which may be called constructive structural realism, is much more complicated (Cao, 1997, 2003a, b; 2006). More discussion on this position will be given in Chapter 9, when the conceptual development from current algebra to QCD is clarified and analyzed. For the present purpose, it suffices to list two of its basic assumptions: (i) the physical world consists of entities that are all structured and/or involved in larger structures; and (ii) entities of any kind can be approached through their internal and external structural properties and relations that are epistemically accessible to us. Its core idea that differentiates it from other versions of structural realism is that the reality of unobservable entity can be inferred from the reality of structure. Methodologically, this suggests a structural approach to unobservable entity, as will be illustrated in the following chapters, and further elaborated in Chapter 9.

On the basis of a structural understanding of unobservable entity and of a dialectical understanding of the relationship between a structure and its components,[3] the third position can also address Kuhn's claim of ontological discontinuity by a notion of ontological synthesis, which underlies a dialectic understanding of scientific development (Cao, 1997, Section 1.4; 2003c). More generally, if a scientific revolution can be understood as a reconstruction of fundamental ontology in the domain under investigation by having reconfigured the expanded set of structural knowledge of the world, then the ontological continuity and progress in scientific development may be understood in terms of reconstructing fundamental ontology in the domain through reconfiguring the expanded set of structural knowledge in a way that is different from the ways in previous theories, and in which empirical laws can be better unified. More discussion on this point will be given in Chapter 9.

Metaphysically, the constructive version differs from the ontic version in having retained a fundamental status for entity ontology, while stressing that this fundamental ontology is historically constructed from available structural knowledge of reality. For this reason, the fundamental ontology of the world has an open texture and thus is revisable with the progress of science. This point and the more general relationship between physics and metaphysics will be examined in Chapter 10.

In addition to exemplifying how successful the structural approach is for discovering unobservable entities, such as quarks and gluons, this enquiry will also shed new light on what has been achieved in the formulation of QCD: it is more than merely a discovery of new entities and forces, but rather a discovery of a deeper level of reality, a new kind of entity, a new category of existence.[4] The enquiry would further help historians of science to understand how such a discovery was actually made through a structural approach. Essentially it takes four steps.

But before elaborating the four steps, let me comment on the structuralist understanding of current algebra. First, without a physical interpretation, a purely mathematical structure, here a Lie algebra, would have no empirical content. Second, if we interpret the Lie algebra in terms of physical structures, taking electromagnetic and weak currents as its representations, then we have physical content, but only at the phenomenological level. In order to understand the physical structures (the currents) properly, we have to move deeper onto the level of their constituents (hadrons or quarks) and their dynamics so that we can have a dynamic understanding of the behavior of the currents, and thus of many features of current algebra and of reasons why current algebra is so successful.

Driven by the recognition of this necessity, most physicists took the idea of quark realistically and tried to conceive it as a new natural kind through the structural knowledge of it.[5] The result was fruitful: there emerged a detailed picture of the microscopic world with quarks as important ingredients.

Now let me turn to the four steps I have just mentioned.

First, the notion of unobservable entities (quarks and gluons) was hypothetically constructed under the constraints of acquired structural knowledge about hadronic phenomena, such as various symmetry properties in hadron spectroscopy and hadronic weak and electromagnetic interactions, which were summarized in the achievements of the current algebra approach to hadron physics. The approach itself was based on the flavor SU(3) symmetry and a hidden assumption, implied by the infinite momentum framework adopted in current algebra calculations, that certain types of interactions among quarks during high energy processes should be ruled out.[6]

Second, the reality of some of the defining structural features of these entities, which were expressed in the current algebra results, was established by checking with experiments, such as the experiments of deep inelastic electron–proton scatterings performed at Stanford Linear Accelerator Center (SLAC). The most important features of quarks and gluons established by the observed scaling in the deep inelastic scattering experiments were their point-like nature and their lack of interactions at short distances.

Third, a coherent conceptual framework, such as QCD, is constructed to accommodate various experimental and theoretical constraints, such as the observed scaling and pion-two gamma decay rate in the former, and infrared singularity and scale anomaly in the latter.

And, finally, the distinctive implications (predictions) of the theory (such as the logarithmic violation of scaling and the three-jet structure in the electron–positron annihilation process) were checked with experiments to establish the full reality of the unobservable entities, quarks and gluons. Although these particles may not be understood as Wigner particles with well-defined spin and mass and new superselection rules based on the liberated color charge, the physical reality of these particles, according to the criteria we will elaborate in Chapter 9, is beyond doubt.

It is clear that the reality of discovered unobservable entities, here quarks and gluons, is highly theory dependent. If the theory, QCD, stands firmly with observations and experiments, the reality of quarks and gluons is confirmed. What if QCD turned out to be wrong tomorrow or in the next decade? More discussions on this interesting question will be given in Chapter 9.

The following chapters will also show that structural realism can help historians of science to make proper judgments on what steps taken were original, consequential, crucial and historically effective in the process of discovery. More elaboration on the light shed by this enquiry on the methodology of historiography will be given in the concluding chapter, but a few remarks on presentism seems to be in order before we embark on a very selective treatment of such an important episode in the history of science.

As a historical enquiry, the aim and real content of historiography of science is to clarify the historical significance of scientific endeavor rather than to popularize its cognitive meaning. Since the perceived meaning and effects of scientific explorations change dramatically with the change of perspectives over time, their long-term significance for science, metaphysics and culture in general that was discerned and understood by contributing scientists when the history was in the making is usually quite different from what critical historians understand because of the difference in perspectives. Thus scientists' judgments, even those concerning their own contributions, cannot be unreflectively taken for granted, but have to be critically assessed and properly interpreted by historians of science before they can be adopted in a historical account.

An elementary but crucial point here worth noticing is that data for historical enquiry are almost always too many and too few. Too many so that we have to select relevant, interesting and informative ones from numerous noises; too few for a meaningful picture of what actually or even plausibly happened in the past so that we have to fill the gap with our reconstructive efforts. How can a historian select events from what are available to him, reconstruct a narrative of what happened in the past, and interpret their historical significance without some guiding hypotheses, that is, without some presuppositions about what had meaningfully happened in the past, how and why an event evolved to the next, and what the overall direction is in the evolution? Thus there is simply no way for a historian to escape from taking working hypotheses and imposing a narrative structure in general, and an overall narrative direction in particular, onto a set of selected events in the past under investigation. It is historians' efforts of this kind that have fixed the meaning structure of a chosen set of past events, from which the past events are interpreted and turned out to be a history through historians' narrative.

But then the important question is where these hypothetic moves of historians come from and what the nature of these moves is. Since the intelligibility and significance of scientific events in the past lie in the message they deliver, the lesson they teach and the authority and confidence they give to current engagements in science and/or culture, any specific hypothetic move must come from a historian's response to imperatives that are immanent in the current praxis in science and/or culture and shaping the prevailing knowledge of the past and expectations for the future. Such a response in fact has defined a historian's intellectual horizon. In this sense, presentism is inescapable in any historical enquiry, and what Croce once declared is right that all history is contemporary history. Similarly, we may say that all historiographical work in science is dictated by a contemporary perspective in science and/or culture, which is particularly chosen by a historian from many investigations, some of which may be in conflict with each other.

Does the acknowledgment of inevitable presentism entail an endorsement of Whig history, a practice in historiography that involves selecting only those data that seem to point in the direction leading to the present without taking proper account of their historical context? Of course, a Whig history is not a real history, but only a distorted retrospection guided by a teleological view of history in terms of a unilinear trajectory. While a Whig history is a form of presentism, the practice of presentism may take other forms, in which the selection of events and the interpretation of their meaning are guided by views of history other than teleology.

A crucial notion whose meaning has to be clarified in this context is that of direction. It is difficult to conceive a narrative without some sense of direction. However, a progressive history or a unilinear direction of events in the past, is too speculative and too apriorist to be acceptable. Even the very notion of an internal direction of events is dubious and thus unacceptable because the direction of events, under the pressure of the contingent circumstances and idiosyncratic strategic considerations of the agents, frequently and unpredictably changes. That is, it is impossible for events in the past to have any coherent direction, and thus nothing is predictable for future developments.

Still, we can legitimately talk about a direction in historical enquiry, that is, a direction in the narrative in which events move toward the end of the narrative (such as the discovery of QCD). However, this direction only reflects the selection of the narrator, who knows the significance of each event in the past by hindsight, and thus has nothing to do with the direction of the events themselves.

It should be noted that the narrator's direction is realized only in frequent changes of the direction of events under investigation. More specifically, as will be illustrated in the following chapters, the direction of scientists' activities changes under the pressure of each stage of scientific exploration: each stage has its own major concerns and means and ways of addressing these concerns; however, sooner or later, the explorations would bring some