Rationality in Flux – formal representations of methodological change

Jonas Nilsson and Sten Lindström
Department of Historical, Philosophical and Religious Studies
Umeå University

A central aim for philosophers of science has been to understand scientific theory change, or more specifically the rationality of theory change. Philosophers and historians of science have suggested that not only theories but also scientific *methods* and *standards* of rational inquiry have changed through the history of science. The topic here is methodological change, and what kind of theory of rational methodological change is appropriate. The modest ambition of this paper is to discuss in what ways results in formal theories of belief revision can throw light on the question of what an appropriate theory of methodological change would look like.

Methodological states

Let us start by introducing the term ”methodological state”. Apart from beliefs, theories and cognitive goals, an agent involved in scientific research has a number of methodological rules or standards of scientific rationality. These standards are of different kinds. Some standards are *heuristic*, prescribing that one should do, or try to do, certain things, such as ”try to find causal explanations for observed phenomena”, ”test theories by making controlled experiments” or ”avoid ad hoc hypotheses”. Other standards are *evaluative*, telling us for example how we should choose between competing theories: ”prefer theories which have been used to make novel predictions over theories which have merely been made to square with the observations made”, ”prefer simpler theories to less simple ones” or ”a successor theory must retain all the corroborated empirical content of its predecessors”. General principles of rationality, such as ”act in such a way that you promote your goals” or ”be prepared to listen to criticism of your beliefs” also function as standards. Logical laws and inference rules may also give rise to standards of rationality, or so it seems. For example: “You ought not to have contradictory beliefs” or “You ought to believe obvious logical consequences of what you believe”.

Furthermore, there are meta-standards pertaining to the evaluation of other standards (or methods), such as ”a method which is more reliable should be preferred to an alternative method which is less reliable” or ”general methods should be coherent with judgments about the rationality of particular episodes in the history of science”.

An agent, which may be an individual or a group of researchers, accepts a large number of normative methods or standards of these different types. All the standards accepted by an agent at a particular time constitute that agent’s methodological state. If an agent comes to accept a new standard, or reject one she previously accepted, she moves from one methodological state to a new such state.

Philosophical theories of methodological change

Whereas some have seen previous changes of scientific standards as pervasive and have tried to formulate models in which all standards are seen as open, in principle, to revision (see Briskman 1977; Laudan 1984 and 1996; Shapere 1984), others have instead argued that such changes as there have been are peripheral and that they can be explained as rational (or irrational) on the basis of some core set of standards that has remained constant (Worrall 1988; Newton-Smith 1981). We shall not try to take a stand on the issue of how extensive previous methodological changes have been, or discuss whether all standards are in principle revisable or if some standards must be treated as immune to revision. Instead we will merely assume that there has been methodological change in science, and that current standards are themselves open to future improvement. The question we want to consider here is what kind of philosophical account should be given of such methodological change.

Say that a certain methodological change is made in some scientific field: An old method is rejected, or a new one is added. If this is rational, there has to be some standard or method, which is used to evaluate the initial methodological state as problematic, and to evaluate the change to a new methodological state as rational. How, then, are standards evaluated? What standards or methods should be applied to determine the rationality of methodological change?

Philosophers of science who discuss methodological change often think of standards in an instrumental way, and their discussions depart from some *axiology*: some conception of the goal or goals of science. This axiology is presupposed as a background when discussing what the appropriate scientific standards are, and how methodological change should be evaluated. Among possible goals that are often mentioned are truth, true explanatory theories, maximizing predictive power, high verisimilitude, empirical adequacy, or problem-solving ability. A usual meta-methodological strategy is then to find some subgoal which is appropriately related to the ultimate goals of science.

Popper, for instance, took something like approach to general, true explanatory theories to be the goal of science, and proposed that a proper subgoal to aim at is to maximize the degree of falsifiability of theories and test them severely (Popper 1989)). Newton-Smith proposes that the ultimate goal of science is theories with high verisimilitude, but that what we must aim at is theories with long-term observational success.

According to Popper and others, the considerations used to select standards (and thus to evaluate methodological change) are broadly logical and philosophical: which standards will ensure that increasingly falsifiable theories are proposed and tested?[1]

According to Newton-Smith, Laudan and others, the considerations are instead empirical: what standards have actually contributed to the selection of theories with such properties as long-term observational success (Newton-Smith 1981) or problem solving efficiency (Laudan 1984, 1996)? Empirical theories of methodological change have been rather popular, and are often associated with naturalistic conceptions of scientific method and methodological change.

Fixed core theories and bootstrap theories

What should a general theory of methodological change look like then? Suppose an agent revises her initial methodological state S_1 in such a way that she enters the new state S_2. In a theory of methodological change, one would like to answer questions like the following: When is it rational to revise a methodological state? And, when is it rational for an agent to make the transition from methodological state S_1 to a new state S_2? In answering these questions, it is relevant to consider two other questions.

(1) Is there a specific *core* of standards (meta-standards) for evaluating methodological change, or is the evaluation of standards and methodological change more varied and pluralistic, so that in principle any method or standard may be relevant to the evaluation of standards? Should a theory of methodological change be a ”*core theory*” or a *pluralistic theory*?

The latter alternative strikes us as more plausible. It seems reasonable that different kinds of considerations – empirical, logical or broadly philosophical – may be relevant for evaluating standards. We think this provides part of the motivation for a bootstrap theory.

[1] This formal approach to the evaluation of methods is defended in Niiniluoto 1999, ch. 6.

(2) The other question is this: Should standards for evaluating methodological change be regarded as necessarily *fixed*, or themselves open to change and improvement? For every pair of methodological states <S_1, S_2>, which is such that the transition from S_1 to S_2 is rational, is there a set of meta-standards according to which all such transitions are rational? That is, should a theory of methodological change be a *static* theory or a *dynamic* theory?

Again, the latter alternative strikes us as more plausible. If improvements in standards of empirical testing is possible, or if improvement of logical and formal standards is possible, it seems reasonable that such improvement could also benefit our resources for evaluating methodological change. This, we think, is a further motivation for exploring bootstrap theories of methodological change.

A *static core theory* would (if fully spelt out) contain a set of standards, which are used to evaluate other standards and changes from one methodological state to another. Outlines of such theories have been sketched by for example Newton-Smith (1981) and Worrall (1982, 1989), and we think that some such theory is presupposed by many philosophers of science who discuss methodological change.

The main motivations for a *bootstrap theory* of methodological change are instead that the evaluation of standards is likely to be a pluralistic matter – in different situations different standards or methods may be applicable – and that a theory of methodological change should itself be dynamic – one does not want to exclude the possibility that the standards used to evaluate methodological change may themselves be improved as science progresses. We shall not try to argue here that a theory of methodological change should take the form of a bootstrap theory rather than a fixed core theory, but rest content with indicating why it is an interesting alternative worthy of further development.

Outline of a bootstrap theory of methodological change

What does a bootstrap theory say? Let us say that as a science develops it goes through not only a sequence of theoretical states, but also a sequence of methodological states. A methodological state is here seen as the set of scientific standards or methods accepted at a certain point in time (in a field or by a group of scientists).

The main bootstrap idea is that some standards in such a methodological state are used to evaluate certain other standards or methods, or the state as a whole, as problematical. Therefore, what particular standards are used for the purpose of evaluating methodological change varies with the particular type of problem detected (say, a logical problem, or empirical evidence suggesting some method is unreliable). Furthermore, what standards are available for evaluating methodological change may change from one stage in scientific inquiry to another.

A bootstrap theory is neutral about which specific standards should be used to evaluate methodological states and methodological change. It is thus compatible with a pluralist view of the evaluation of methodological change, and with a dynamic view of standards for methodological change.

The bootstrap idea is instead to lay down requirements for how standards accepted at a particular point in time (making up a methodological state) may be used to evaluate other standards or a methodological state as a whole, as well as transitions between such states. These requirements we call ”bootstrap standards”.

What is it that drives methodological change according to a bootstrap theory? Well, it may be different kinds of input which motivates scientists to revise their standards. The impetus may come from empirical information about the track record of some method which constitutes evidence that it is unreliable, or it may be new logical or philosophical arguments, or a perception of disequilibrium within a methodological state.

Versions of bootstrap theories of rationality have been proposed earlier by Briskman and Laudan. An early bootstrap theory of methodological change was proposed by Briskman already in 1977. His main idea is that in research certain kinds of problems arise (“problems of preference” and “problems of goal-pursuit”) which cannot be solved by using existing methods or standards. A methodological change is rational to the extent that it solves such problems. The problems encountered function as standards for evaluating methodological changes.

Laudan has also proposed a bootstrap theory.[2] The central idea is that standards are to be seen as means for achieving scientific goals, and that standards and methodological changes can be evaluated in terms of how efficient they are as means for achieving the goals of scientific research.

In his dissertation Nilsson argued that previous bootstrap theories failed to account for the details of the bootstrap processes where standards are changed, and in a later paper (Nilsson 2005) he proposed a general bootstrap theory. In distinction from previous theories it is explicitly formulated in terms of how methods or standards operative at one scientific stage can be used to evaluate methodological change at that stage. The theory contains a number of bootstrap standards, which are held to govern rational changes of method. To illustrate the

[2] In Laudan 1984 it is called ”the reticulated model of scientific rationality” whereas in Laudan 1996 it is called ”normative naturalism”.

contents of such a theory, here is a tentative list of informal bootstrap standards Nilsson proposed (Nilsson 2005).

> Suppose an agent or group of agents accept a set of standards S_1 and revise some of these so that they come to accept a new methodological state S_2. For the transition from S_1 to S_2 to be rational, the following requirements should be met:
>
> *Conservatism*: It is rational to revise a methodological state S_1 only if there is some reason to regard S_1, or some part of S_1, as problematic.
> *Internal Conformance*: The standards used to evaluate S_1 or part of S_1 as problematic must themselves be part of S_1 (they must be standards accepted by the agent).
> *Problem Solving*: The particular problem identified in S_1 must be absent from S_2.
> *Stability*: S_2 must be better than S_1 according to the standards in S_2.
> *Prospective Acceptability*: S_2 must be better than S_1, according to the standards that are members of S_1, except for those standards in S_1 that are criticized and revised.
> *Goal-pursuit:* A change from S_1 to S_2 must not be such that it is judged to become more difficult – according to the standards in S_2 and those standards in S_1 that are not being criticized and revised – to achieve the scientific goals operative at that point in time.

The bootstrap theory presented in Nilsson 2005 simplifies matters in an important respect: it treats the standards accepted by an agent – a methodological state – as a pure set and specifies how one part of that set can be used for evaluating another subset of standards. It does not take account of the different kinds of relations that hold between the different standards, thus treating methodological states as unstructured.

The further development of the theory would consist partly in describing these relations and formulating bootstrap standards, which prescribe how such relations are relevant to the rationality of methodological changes. For this purpose, constructing models of sets of standards or methods is likely to be fruitful as it may make it easier to discern and investigate patterns of relations holding between standards.

Should a bootstrap theory be formulated in such a way that the bootstrap standards belong to a metalevel which is separated from the object level of other standards? Philosophically it seems natural instead to formulate a bootstrap theory as a *one-level* theory. That would mean that the bootstrap standards themselves function on the object level, within the methodological states themselves.

When it comes to the question of how theories of methodological change should be formally represented, two questions arise in particular for bootstrap theories: Are there problems of formally representing bootstrap standards, over and above problems with representing other standards that can be applied to methodological change? And, are there obstacles to formally representing a bootstrap theory as a one-level theory?

We hope that bringing mathematical and other formal tools to bear on methodological states will make it easier to uncover and theorize about interesting structural features of such states. The mathematical models in question may be constructed along the lines of the BDI-model of rational agency.

The BDI-model of rational agency

In this part we will discuss the possibility of studying methodological change from a formal or logical point of view. We start out by briefly describing some of the work that has been done in philosophy and artificial intelligence (AI) concerning the architecture of rational agents; and the cognitive dynamics of such agents. Much of the work has of course been concerned with the logic of belief change (belief revision and belief update), but researchers in AI have also created models of the dynamics of rational agents with goals, intentions, plans, etc. and the ability to act. The development of such agents is governed by very general laws of practical reasoning, roughly: If an agent has certain beliefs and certain goals, then he chooses some available course of action that he believes will favour his goals. A rational agent modifies his beliefs about the world on the basis of the information he receives. And he modifies his immediate goals (intentions) accordingly as his beliefs change. Thus AI researchers have not only studied rational belief change but also rational changes in goals, intentions and plans.

Here we want to discuss the possibility of adapting and extending the kind of models of rational attitude change developed within philosophy and AI to the modelling of rational change within science. The basic idea is to view a scientific research community as an agent with beliefs, goals, procedures, etc. We are not going to consider the interaction and communication between the members of such a community. In reality a research community may be far from homogeneous; there may be differences in opinions and goals between its members and it may be of great interest to study the dynamics within such groups. It is presumably also of great interest to study how different research communities with quite different research programmes may communicate and influence each other. Here, we will make, the no doubt, severe idealization that research communities can be treated as single agents that do not interact with other research communities.

Another question that we do not discuss is the one concerning the principles of individuation of agents in general, and research communities (or research traditions) in particular. In our special case, when is it correct to say that a community observed at time t is the very same research community as one that we observe at a later time t'? Presumably there has to be a continuous development tying the two stages together in order to say that they are stages in the development of one research community (or belong to the same tradition). A question that may be even more fundamental is also ignored: what kinds of entities can be rational agents? Within AI the conception of a rational agent seems to be quite liberal: humans, robots, even entities living in "virtual reality" are described as being rational. Philosophers are usually more restrictive. We are only assuming that collectives of humans, in particular societies of researchers may be described as having beliefs, goals and plans, and being rational or irrational.

The BDI-model is an architecture for constructing software for intelligent machines inspired by the belief-desire-intention theory of human practical reasoning developed by Michael Bratman (Bratman 1987, 1999). According to this model an agent has at a given time a set ***B*** of *beliefs* and a set ***G*** of *goals* (or desires). The agent's beliefs correspond to information that she has about the world. We assume that the belief set ***B*** is a consistent set of propositions. ***G*** is a set of propositions representing states of affairs that the agent would like to see realized. We do not assume that ***G*** is consistent: the agent may very well have contradictory or opposing goals that cannot be realized simultaneously. However, there is at any given time a subset ***I*** of the agent's goals that she is committed to realizing. These are the agent's *intentions*. The set ***I*** is assumed to be consistent. At any time the agent's intentions are determined by her beliefs and goals at that time. The agent's intentions at any given time are the goals that are operational at that time in determining her actions. A natural assumption is that an agent gives up an intention only if she either believes that the intention has been achieved or that it cannot be achieved (with too much effort).

According to the BDI-model, the dynamics of a rational agent may be described as follows: Initially, the agent is in a certain mental state with beliefs ***B***, long term goals ***G***, intentions ***I***, and an active plan ***P*** for realizing her current intentions. Then the agent receives some new information or goes through some process of reasoning resulting in a new belief state ***B'.*** The change in beliefs in turn leads the agent to reconsider her intentions. She then devises a plan ***P'*** for realizing the new intentions in light of her new beliefs, and so on.

Models of rational methodological change

We may think of a scientific research program along the lines of the BDI-model. The agent is now a scientific research community:

Agent: A research community

Beliefs: A scientific corpus consisting of a theory, auxiliary hypotheses, data.

Goals: True explanatory theories, verisimilitude, empirical adequacy etc.

Intention: To test a certain hypothesis (research agenda)

Plan: To perform a series of experiments according to a well-established methodology.

Action: The tests are performed and the results are evaluated.

The results of the tests may then lead to changes in the corpus as well as in the research agenda and the methodological rules. Certain long-term goals may be constitutive of the scientific endeavour. Moreover certain structural (or logical) features, like the general BDI-model may also be characteristic of science. Perhaps one can speak a little vaguely of a logic of scientific reasoning, perhaps open to refinement and revision. However, in accordance with the bootstrap theory of rational scientific change, there are no theories, goals or methods of science that are beyond rational criticism.

Concluding discussion

So what does it mean that a scientific agent accepts certain standards of rationality and what kind of entities are these standards? One idea that needs to be pursued is that accepting a standard of rationality is a mental state (a propositional attitude), namely a certain kind of belief about what we rationally-ought-to believe or do. Hence, on this view, rationality standards are *requirements* of rationality in the sense of Broome (2007). If we prefer to speak in terms of what we rationally-ought-to-do or rationally-ought-to-believe instead, rationality standards are beliefs about what we under the circumstances rationally-ought-to-do or rationally-ought-to-believe. For example, we may believe that rationality requires of us that our beliefs are logically consistent, or we may believe that rationality requires of us that we believe the (obvious) consequences of what we believe, or intend the necessary means for achieving our goals. If so, then these requirements are among the standards of rationality that we accept. Our rationality standards may, of course, also include beliefs about how we rationally-ought-to change our beliefs when we receive new information.

As has been pointed out by Broome and others, a belief that we rationally-ought-to *F*, need not be normative in the strong sense of entailing a belief that we, everything considered, ought to F. Rationality (as we conceive of it) may require of us that we F, although it is not the case that, everything considered, we ought to F. If there are objectively correct standards of rationality, then we may also be mistaken about what rationality requires of us.

If rationality standards are viewed as beliefs about what rationality requires of us, then scientists may deviate from their standards in their actual practice of science. It is natural to think that one function of our standards is precisely to enable us to criticize and correct our scientific practice. On the other hand, it appears that the direction of criticism could under some circumstances be reversed: if a certain scientific practice which we judge to be generally successful fails to meet our rationality standards, then at least prima facie that might constitute a case for considering the rationality standards themselves to be problematic.

Now, if standards of rationality are —or can be viewed as—beliefs of a certain kind, then the theory of methodological change becomes a special branch of a generalized theory of belief change. The formal methods of belief revision theory can then also be applied to methodological change. However, the standard AGM-axioms of belief revision (Cf. Alchourrón, C., Gärdenfors P., and Makinson D.,1985) are not applicable without restriction to methodological change. For example, AGM-revision satisfies *Preservation:*

If A is consistent with the theory $\boldsymbol{T}$, then $\boldsymbol{T} \subseteq \boldsymbol{T}*\mathrm{A}$,

where $\boldsymbol{T}*\mathrm{A}$ is the revision of the theory $\boldsymbol{T}$ with the statement A. However, let A be the statement ”One ought to look out for dodos”. Someone who does not know whether or not there are any dodos around may accept A, although he would give up the belief in A once he learned that the dodo is extinct. Hence, in the presence of deontic beliefs Preservation has to be abandoned.

In an extension of the BDI-model, which includes an agent’s methodological states, the agent is situated in an environment (the external world). The agent has a total (internal) state consisting of (at least) the following components: A theoretical state $\boldsymbol{T}$ (the agent’s current scientific theory about the world), a goal state $\boldsymbol{G}$, certain intentions $\boldsymbol{I}$ for action, and a methodologicial state $\boldsymbol{M}$. Moreover, there is for each total state S a preference relation $\leq_S$ over total states. $S_1 \leq_S S_2$ means that the state S_2 better satisfies the goals and standards that hold in $\boldsymbol{S}$ than does $\boldsymbol{S}_1$. $\boldsymbol{S}$ is in all likelihood not optimal from its own perspective. This fact will move the agent to a state $\boldsymbol{S}'$ that is better than $\boldsymbol{S}$ from the perspective of the current state $\boldsymbol{S}$. The question arises: Can we formulate any informative constraints on this process?

We can distinguish at least four different kinds of change:

(i) Changes in scientific theory in response to new research results.

(ii) Changes in actual scientific practice in order to make this practice conform better to current rationality standards (our beliefs about correct methodology).

(iii) Changes of current rationality standards as a result of a critical discussion of their appropriateness.

(iv) Changes of basic scientific goals and values.

It is changes of types (iii) and (iv) that primarily interest us here. Generally, inquiring agents will prefer to change their theories about the world rather than their rationality standards and their basic scientific goals. Under what circumstances is it instead rational to change, e.g., one's rationality standards rather than one's theories or one's actual practice?

The bootstrap theory discussed above is one attempt to answer this question, by proposing constraints on how different methodological states in a sequence of changes should be related to each other if the process is to be one of rational methodological change. One challenge is then to develop a suitable formal framework, with a language which allows one to represent rationality standards (including meta-standards such as the bootstrap standards) as well as theories, goals, intentions, plans and cognitive actions. A related challenge is to extend the BDI-model of rational agency in such a way that it also covers those rare but interesting occasions when inquiring agents come to the conclusion that what rationality requires of them is to reevaluate their beliefs about what rationality amounts to.

References

Alchourrón, C., Gärdenfors P., and Makinson D., (1985) "On the logic of theory change: partial meet contraction and revision functions", *The Journal of Symbolic Logic* 50, pp. 510–530.

Bratman, M. (1987) *Intentions, Plans, and Practical Reason.* Cambridge, Mass. ; Harvard University Press.

Bratman, M. (1999) *Faces of Intention : Selected Essays on Intention and Agency.* Cambridge: Cambridge University Press.

Briskman, L. (1977) "Historicist Relativism and Bootstrap Rationality". *The Monist* 60, pp. 509-39.

Broome, J. (2007) "Requirements", in *Hommage à Wlodek: Philosophical Papers Dedicated to Wlodek Rabinowicz*, edited by Toni Rønnow-Rasmussen, Björn Petersson, Jonas Josefsson and Dan Egonsson. Lund: Lund University.

Laudan, L. (1977) *Progress and Its Problems: Towards a Theory of Scientific Growth.* London: Routledge and Kegan Paul.

Laudan, L. (1984) *Science and Values: The Aims of Science and Their Role in Scientific Debate*. Berkeley, CA: University of California Press.

Laudan, L. (1996) *Beyond Positivism and Relativism: Theory, Method, and Evidence*. Oxford: Westview Press.

Newton-Smith, W. (1981) *The Rationality of Science*. London: Routledge and Kegan Paul.

Niiniluoto, I. (1999) *Critical Scientific Realism*. Oxford: Oxford University Press.

Nilsson, J. (2000) *Rationality in Inquiry: On the Revisability of Cognitive Standards*. Ph.D. dissertation, Umeå University.

Nilsson, J. (2005) ”A Bootstrap Theory of Rationality”. *Theoria* 71, pp. 182-99.

Popper, K.R. (1989) *Conjectures and Refutations: The Growth of Scientific Knowledge*, 5th revised and corrected ed. London: Routledge.

Popper, K.R. (1980) *The Logic of Scientific Discovery*, 4th ed. London: Routledge.

Worrall, J. (1982) “Broken Bootstraps”. *Erkenntnis* 18, pp. 105-30.

Worrall, J. (1988) ”The Value of a Fixed Methodology”. *British Journal for the Philosophy of Science* 39, pp. 263-75.

Worrall, J. (1989) ”Fix it and be Damned: A Reply to Laudan”. *British Journal for the Philosophy of Science* 40, pp. 376-88.