Max Albert

Critical Rationalism and Scientific Competition*

Abstract: This paper considers critical rationalism under an institutional perspective. It argues that a methodology must be incentive compatible in order to prevail in scientific competition. As shown by a formal game-theoretic model of scientific competition, incentive compatibility requires quality standards that are hereditary: using high-quality research as an input must increase a researcher's chances to produce high-quality output. Critical rationalism is incentive compatible because of the way it deals with the Duhem-Quine problem. An example from experimental economics illustrates the relevance of the arguments.

1. Introduction

In this paper, I consider Karl Popper's critical rationalism under an institutional perspective.[1] Such a perspective is already apparent in Popper's own works (Jarvie 2001) and has been stressed and developed especially by Hans Albert (e.g. 1985[1968], ch. 2, section 6; 2006; 2010). Following this line of thought, the present paper consider methodologies in general, and critical rationalism specifically, in the context of a formal game-theoretic model of scientific competition where the relevant methodology serves as a constitution of science.

This extended institutional perspective offers new arguments that can be used to criticize methodologies. The arguments are based on the fact that a scientific field's methodology generates specific incentives and disincentives for researchers.

Science, like a market economy, is characterized by competition and cooperation.[2] Although science is very different from a market, the two institution are

---

[1] Any consistent interpretation of Popper's own writings must, it seems, be a selective interpretation because Popper dithered between two different ideas. These two different ideas have developed into two versions of critical rationalism, Alan Musgrave's (1993; 1999, ch. 16) 'positive' and David Miller's (1994) 'negative' critical rationalism. Subsequently, I focus on positive critical rationalism.

[2] The term 'science' is used in the European sense here and, therefore, includes the social sciences and parts of the humanities. Of course, methodologies differ between these fields.

similar in one important aspect: both use competition to organize the division of labor.

Researchers in a field compete for attention, status, and material rewards. Nevertheless, competition unfolds within rules that regulate competition, namely, methodological rules. Moreover, researchers cooperate in the sense that they publish ideas and results that are taken up and developed further by their peers.[3] Scientific progress, then, is the result not only of competition but also of an extensive division of scientific labor. Both, competition and cooperation, are taking place under, and are influenced by, the prevailing methodology.

A field's methodology is a constitution, but it is neither designed nor enforced by some central agency. A methodology is developed and enforced in the same decentralized and competitive process it regulates. Not just any methodology can establish itself in such a process. In game-theoretic terms, the methodology must constitute a Nash equilibrium, that is, it must be in the own best interest of a researcher to stick to the methodological rules if he expects other researchers to do so. Or, in the jargon of institutional economics, the methodology must be incentive compatible.

A methodology that is not incentive compatible, whatever its epistemic virtues, cannot be expected to prevail in scientific competition. But even an incentive compatible methodology may not foster cooperation in science. The progress of science we observe would be impossible if the prevailing methodology would reward researchers for ignoring each others' work. A good methodology, then, fulfills three conditions. It is satisfactory from an epistemic point of view, it is incentive compatible, and it induces cooperation in the sense of a division of scientific labor.

The paper proceeds as follows. *Section 2* considers the role of methodology in scientific competition and presents a simple game-theoretic model of scientific competition. The model's equilibrium is characterized by an incentive compatible methodology. The methodology can establish itself because high quality is hereditary in the scientific production process: using high-quality inputs makes it more likely to produce high-quality output. *Section 3* argues that critical rationalism fits the model because of the way it solves the Duhem-Quine problem. In addition, a more general solution is proposed. The paper concludes with a summary of the main points.

2. Methodology and Scientific Competition

2.1 The Problem of Quality Standards

By science as an institution, I mean academic or open science, that is, the whole system of research-oriented universities, scientific journals, the peer review system, learned societies, and so forth. Open science is to be distinguished from

---

[3] Researchers also build cooperative teams that work and publish as a team. Even though this form of cooperation is often seen today, it is of secondary importance. Science would be cooperative even if every researcher worked and published alone.

proprietary science, that is, research under the protection of intellectual property rights like patents.[4]

In open science, the production of high-quality research is the way to rise to the top. Researchers compete for scarce goods: attention, status, and material rewards. These goods are tied to academic positions, research grants, journal space, scientific prizes, and so on. There are incentives to aim at high quality but, of course, no sure-fire method to produce it, which explains why we also observe low quality.

High-quality contributions are rewarded. But there is no central agency deciding on the standards according to which contributions are to be judged. The evaluation rules that are used to distinguish between high quality and low quality are part of scientific methodologies.[5] Depending on the methodology, high quality ideas or results may be those that are tentatively accepted as true, assigned a high probability, or evaluated positively in some other way.

By and large, competing researchers in a field accept common methodological standards. In most cases, these standards are demanding in the sense that high quality is not easy to produce—otherwise, the quality standards could not be used to distribute scarce rewards. How can we explain the existence of common standards?

In markets, product quality is mostly a matter of demand and supply. Consumers get higher quality if they are prepared to pay for it. Such an explanation cannot work in the case of open science. Contrary to popular assumptions, open science is not a 'market of ideas'.[6] It uses the so-called voluntary contribution mechanism, not the price mechanism. Researchers publish their ideas and results, which can then be used free of charge by anybody who wishes to do so. Research outputs are non-rival goods anyway; publication turns them into public goods. Since researchers are typically not paid for their publications but receive a fixed income, this is a case of voluntary provision of public goods.

The voluntary contribution mechanism can be combined with different kinds of incentives. In science, status among one's peers is an important reward in itself and the key to most other rewards, like attractive positions or Nobel prizes. The status of a researcher is determined by his impact on the field, that is, by the extent to which his ideas and results are used by other researchers.[7] All researchers face the same incentives when selecting the ideas and results on which to build their own work. Everybody tries to produce inputs for the next step in research, in a potentially unending chain. From the internal perspective, all outputs are either discarded or used as intermediate inputs in further research.

---

[4] See Diamond 2008 and Stephan 1996; forthcoming on the economic analysis of science and, specifically, on the characteristics of open science. See Albert 2008a for a compact summary and statement of the position underlying the present paper.

[5] See Gadenne 2005, section 1, on evaluation rules in contrast to procedural rules like, e.g., the rules of proper experimentation.

[6] See Albert 2008a. Cf. Vanberg 2010 for a comprehensive discussion of the science-as-a-market analogy from a constitutional-choice perspective.

[7] On incentives for voluntary contributions, see Hackl et al. 2005; 2007. On status as a reward in itself, see Marmot 2004. See Merton 1973 on status as an incentive in science and Hull 1988, 283 and elsewhere, on use as the basis of status. See Congleton 1989 for a model of status seeking and competing voluntary contribution mechanisms.

There are, of course, users of research outside science: innovators, administrators, politicians, doctors, lawyers, and writers of popular science books. However, their preferences do not influence competition within science. Indeed, this is the whole point of scientific competition. It is difficult, if not impossible, for outsiders to evaluate the quality of research, especially cutting-edge basic research. Scientific competition delegates the evaluation to those who have the required competence: the researchers themselves.[8]

Science is therefore characterized by a high degree of autonomy—by producer sovereignty, not consumer sovereignty.[9] Nevertheless, producer sovereignty in science works surprisingly well for the outside users of research, at least on average. The reason is that competing researchers often coordinate on reasonable methodological standards and maintain them collectively. This brings us back to the central question: Why and how do they do it?

In this paper, I consider a very simple mechanism that can explain coordination on common methodological standards. This mechanism is based on a production function for research with a simple but intuitively plausible property: input quality affects, at least stochastically, output quality. In other words, quality is hereditary in scientific production. If the costs of selecting the right kind of quality are not too high, hereditariness implies that the expectation that everybody prefers inputs of a certain quality is self-fulfilling. Since everybody produces inputs for other researchers, it is in everybody's own interest to build upon research of the required quality, in order to maximize the chance to produce output of this quality. Under these assumptions, methodologies are incentive compatible: if everybody adheres to the methodological rules, it is in everybody's interest to do so.

A link between input and output quality provides a rationale for researchers to be selective in the choice of inputs. Researchers use the work of others if they think that it is good enough to build upon it (Hull 1988). Hence, if a paper gets a negative evaluation, it will not be used by other researchers, which means that the author fails in the quest for status. Thus, given hereditary quality, high-quality papers are used and low-quality papers are discarded in equilibrium just because everybody expects this kind of behavior.[10] This also means that evaluation rules are revealed in the choices made by researchers when they decide which ideas and results to use in their own research.

---

[8] Cf. the theoretical and historical explanation of the self-regulating character of open science offered by Dasgupta/David 1994 and David 1998; 2004 which centers on the inability of the patrons of science to evaluate its results. This explanation must be supplemented, however, by a model of scientific competition explaining how self-regulation can actually deliver high quality. This explanation problem becomes even more severe once it is recognized that scientific quality standards are endogenous to science.

[9] The term producer sovereignty in this context is due to Mayer 1993, 10, who considers it to be a problematic feature, at least in economics (see also Vanberg 2010, 45).

[10] A similar process can lead to the adoption of common standards in the production of durable consumer goods, especially houses, see House/Ozdenoren 2008.

2.2 Hereditary Quality and the Game of Science

Let me sketch the mechanism of hereditary quality in the simplest terms.[11] Consider an infinite sequence of researchers. Researcher 0 publishes a paper. Researcher $t = 1, \ldots, \infty$ has two options: publishing a completely new paper or building upon the paper of his immediate predecessor, researcher $t - 1$.

All researchers have identical utility functions. They receive a payoff of v_1 if their paper is used by their immediate successor, and $v_0 < v_1$ if their paper is not used. This reflects, in the simplest way possible, the idea that researchers gain status, and other rewards tied to status, if they have an impact on their field. Research has costs depending on what the researcher tries to achieve and whether he can use the paper of his predecessor or not. If the costs of research are c, the overall payoff of researcher $t > 0$ is $u(v, c) = v - c$, where c and $v \in \{v_0, v_1\}$ depend on the researcher's own decision and the decision of his immediate successor, respectively.

We now introduce different qualities. Let a paper fall into one of two categories, X or Y. For instance, the ideas in a paper may be either simple or complicated, or either consistent or inconsistent. Writing X-papers is the more demanding task. We assume that researchers can tell an X-paper from a Y-paper, but we do not assume that the category of their own or any other paper enters their utility function.

The costs of research have several components. There is a basic cost that is independent from the researcher's strategy; we assume that this cost is already included in the payoffs v_0 and v_1. If a researcher discards the paper of his predecessor and starts from scratch, he bears an additional cost of c_D. If a research tries to write an X-paper, which is more demanding, he bears an additional cost of c_X.

The core assumption of the model is a production function introducing hereditary quality. If a researcher wants to produce a Y-paper, success is certain. Producing X-papers is more demanding; success is uncertain. The probability of producing an X-paper is p if one builds upon an X-paper, r if one builds upon a Y-paper, and q if one uses no paper and starts from scratch, with $1 > p > q > r > 0$.

This simple production function assumes that researchers are all alike in their abilities. It may be the case that not everybody is able to become a researcher; but those who do enter research have identical abilities and preferences.

We assume perfect information. All researchers know all the other researchers' costs and payoffs, and researcher $t > 0$ knows what all researchers $s < t$ did and whether they produced an X-paper or a Y-paper. The game, then, is a sequential game. This game has many equilibria. However, with the help of some plausible requirements, we can restrict considerations to two equilibria.

---

[11] For full-fledged models along the lines indicated here, see Albert 2006; 2008b. For an extended discussion and interpretation of the present simple model, see Albert 2011. For a commentary on the present approach to methodology and economic approaches in general, see Strevens 2011.

We consider only subgame perfect, strict, and stationary equilibria in forward-looking strategies, subsequently called 'simple equilibria' for short.

Subgame perfection is a standard assumption that ensures that equilibria are not based on threats or promises that a player would not carry out if the relevant situation arose.

Strictness means that a player would actually lose by unilateral deviation from the equilibrium. Strict equilibria are self-fulfilling prophecies; in a non-strict equilibrium, at least one player is indifferent between playing his equilibrium strategy or some other strategy.

Stationary equilibria are equilibria where all researchers—with the exception of the first, who lacks a predecessor whose paper he could use—use the same strategy. A stationary equilibrium describes a situation where quality standards are established, nobody expects methodological change, and all researchers comply with the accepted standards. Non-stationary equilibria would describe situations where methodological change unfolds according to everybody's expectations. This seems implausible. Of course, methodological rules change over time. However, these changes seem to be unexpected. They should accordingly be modeled as unexpected changes from one equilibrium to another.

Forward-looking strategies are strategies that may require different actions depending on the quality of the predecessor's paper but, given this quality, will ignore all past events. Thus, punishment strategies and trigger strategies are ruled out, which is reasonable in this context since these strategies do not seem to play any role in science.

There are just two simple equilibria (see Albert 2011, appendix C): a non-discriminating equilibrium where all researchers (except the first) use their predecessor's paper and write Y-papers, and a discriminating equilibrium where only X-papers are used and every researcher tries to write an X-paper.

In both equilibria, researchers do, in their own best interest, what they expect their successors to do. In the non-discriminating equilibrium, each researcher expects his successor to use his paper no matter whether it is an X- or a Y-paper. Therefore, it makes no sense to do anything else but to produce a paper as cheaply as possible, which means using his predecessor's paper and writing a Y-paper. Obviously, the non-discriminating equilibrium always exists under our assumptions.

In the discriminating equilibrium, each researcher knows that his successor will only use X-papers and, therefore, tries to write such a paper. He will even discard his predecessor's paper if it is a Y-paper and start from scratch in order to maximize his own chances to produce an X-paper. However, a discriminating equilibrium may not exist if the costs of pursuing this strategy are too high in view of the expected payoffs. Existence of the discriminating equilibrium requires $v_1 - v_0 \geq \max\left\{\frac{c_D + c_X}{q}, \frac{c_D}{q-r}\right\}$.

The expected utility of a researcher in the non-discriminating equilibrium, EU_N, is, of course, maximal since he bears only the basic costs of research and his paper is always used: $EU_N = v_1$. The expected utility of a researcher in the discriminating equilibrium, EU_D, must be lower for two reasons: he bears

additional costs, and he cannot be sure that his paper will be used. On average, the expected welfare loss for researchers through the quality standard is $EU_N - EU_D = c_X + (1 - \chi)(v_1 - v_0 + c_D)$, where $\chi = \frac{q}{q+1-p}$ is the long-run frequency of X-papers (see Albert 2011, appendix D). This is easy to interpret: c_X is the cost of trying to produce an X-paper; $v_1 - v_0 + c_D$ is the cost that results from discarding a Y-paper, namely, $v_1 - v_0$ for the discarded paper's author and c_D for his successor.

The model shows that demanding standards can be established in scientific competition even if researchers are disinterested in them. The expectation that others will use only work satisfying the standard is self-fulfilling. The model does not show, however, how such a self-fulfilling expectation may arise; in fact, the model also allows for an equilibrium where the standard is not established. Whether a standard is established or not is, in practice, a matter of methodological discussions, which create expectations about the standards research must live up to in order to be taken seriously.

3. Critical Rationalism

In this section, I argue that critical rationalism is an incentive compatible methodology that induces cooperation in the sense of a division of scientific labor (cf. also Albert 2002; 2004; 2007). First, I provide an up-to-date sketch of critical rationalism. Second, I present an example from experimental economics, the ultimatum experiment, which shows critical rationalism in action. Third, I relate critical rationalism to the mechanism of *section 2*'s model, using the ultimatum experiment as an illustration of the argument.

3.1 Critical Rationalism as a Decision Theory

Critical rationalism is based on the belief that the critical method is the best way to find out the truth about things. The critical method can partly be described by decisions rules stating under which conditions to believe or accept statements if one aims at true beliefs. I say, 'partly' because the decision what to believe or accept does not exhaust the critical method.

Decision rules for the acceptance of statements are evaluation rules: they evaluate statements by assigning them truth values, if in a tentative and fallible way. Here is a rough version of the basic decision rule of critical rationalism.[12]

---

[12] The following account is an interpretation of Musgrave's (1993; 1999, ch. 16) positive critical rationalism. See also Andersson 1984; 1994; Gadenne 2006 and, critical, Miller 1994. The presentation follows Albert (forthcoming) in separating the epistemology of critical rationalism from the definition of rationality, which is irrelevant here.

Critical Acceptance (CA) Rule: Accept a statement if and only if falls into one of the following categories:

1. It describes an observation and has not succumbed to criticism.

2. It is a non-observational statement that is well-corroborated and has no equally well-corroborated alternative.

3. It follows deductively from statements in the first two categories.

Ad 1: The possibility to criticize observation statements is the main point of the solution to Popper's problem of the empirical basis (cf. Andersson 1984; Musgrave 1999, ch. 16). One possibility is to make explicit the implicit theoretical content of an observation statement, for instance, the implicit assumption that the conditions under which the observation was made allowed the observer to observe correctly. When this implicit assumption is stated in the form of a general hypothesis, it can be tested. If the hypothesis is falsified, the original observation statement can be rejected.

Ad 2: A non-observational statement is corroborated if it has survived serious criticism. Serious criticism in the case of nomological statements means serious attempts at falsification, that is, severe tests. This is the falsificationist core of critical rationalism.

In the case of metaphysical statements, serious criticism may just mean serious philosophical discussion (see Musgrave 1999, ch. 16). A metaphysical statement is viewed as corroborated if it comes out better than the alternatives proposed so far, possibly because the alternatives share all its disadvantages but not all its advantages.

Ad 3: This rule is so widely accepted that it need not be discussed in the present context.

Critical rationalism is not a normative approach. The CA rule just describes an important part of the critical method in terms of rules. The reason why critical rationalists use the CA rule is that they accept a statement like the following one.[13]

Critical Rationalism (CR): If one wants to believe statements if and only if they are true, the best decision rule is the CA rule.

'Best decision rule' means 'best among the known options'—like, for instance, induction, Bayesianism, and others. But even the best decision rule cannot be a reliable decision rule; the CA rule, and the critical method in general, cannot avoid error. Possibly, the CA rule is best because the alternatives are useless. For instance, Bayesianism, the modern probabilistic version of induction, is empty (Albert 2003; 2009).

---

[13] I have separated the CA rule from CR only for convenience. We could easily work with a single positive stament along the following lines: 'If one wants to believe statements if and only if they are true, the best option is to believe statements if and only if they fall into one of the following three categories ...' Obviously, neither categorical nor hypothetical imperatives nor any other normative statements are involved. Critical rationalism is a technology (cf. Hans Albert 1985[1968], ch. 2, section 6; Gadenne 2006, section 1).

The belief in CR is not dogmatic. If Musgrave (1999, ch. 16) is right, CR itself, which is metaphysical, can be accepted according to part 2 of the CR rule: the CA rule itself survives critical discussion, emerging as the best method of belief formation if one aims at true beliefs. This need not be the case. Nothing in the CA rule ensures that CR survives a critical discussion. Indeed, those who reject CR argue that it does *not* survive critical discussion. Even its opponents, then, must agree that the CA rule does not automatically lead to the acceptance of CR.

The CA rule is not a complete description of the critical method. It just solves the problem of deciding whether, under given circumstances, a given statement should be believed or not. However, it does not say which statements should be picked out for critical scrutiny. Since it is not possible to scrutinize all one's beliefs at once, the CA rule can only be used locally and sequentially.

In the rest of the paper, the focus is on empirically testable scientific hypotheses. According to critical rationalism, the most important quality of such a hypothesis is that it has survived severe tests, that is, tests that could have led to a falsification. This is the falsificationist core of the CA rule. Hypotheses that have survived severe tests are called corroborated.

Hypotheses, then, fall into one of three categories: untested, falsified, or corroborated.[14] They are categorized and re-categorized, very roughly, as follows. A new hypothesis is born untested. If it passes severe tests, it turns into a corroborated hypothesis. If it fails a severe test, even a well-corroborated hypothesis becomes falsified. And even a falsified theory may be restored to its former status if the falsifying observation statements are successfully criticized.[15]

3.2 The Case of the Ultimatum Experiment

Experimental economics provides us with many examples of critical rationalism at work. We take a look at a prominent case, the ultimatum experiment (Güth et al. 1982; Camerer 2003, ch. 2).

In the ultimatum experiment, two participants divide a sum of money provided by the experimenter. Each of the two participants is assigned one of two roles, proposer or responder. The proposer makes an offer, like '60% for me, 40% for you'. The proposal is an ultimatum; the responder can only accept or reject the offer. If he accepts, the proposal is implemented; otherwise, the experimenter takes his money back, and both participants get nothing. These rules are explained to both participants in the instructions.

We consider this experiment as a test of the homo oeconomicus (HO) model, that is, the nomological hypothesis that people are rational, egoistic, and moti-

---

[14] Miller's (1994) negative critical rationalism seems to require that untested and corroborated hypotheses are treated in the same way, so that it would suffice to distinguish only two categories, falsified and non-falsified. However, the different treatment of untested and corroborated hypotheses is necessary for incentive compatibility; see below.

[15] On corroboration and severe tests, see Gadenne 1998a and the exchanges in Mayo/Spanos 2010, esp. Musgrave 2010 and Mayo 2010. For more on severity, see 3.3 below. In 3.4 below, it is argued that not all hypotheses are born equal.

vated by material interests alone.[16] Of course, the HO model had been falsified
before, for instance, in experiments demonstrating preference reversals (see Roth
1995, 65–72 for an overview). However, it would be a mistake to assume that
one falsification of an important hypothesis would make further falsifications
irrelevant. The history of experimental economics shows that it is extremely
important to explore where and how the predictions deriving from a hypothesis
break down.

Let me put this slightly differently. Even if the HO hypothesis is false, it may
still be the case that people are rational, egoistic, and motivated by material
interests alone in simple strategic situations—situations that are simple enough
so that preference reversals and other failures of rationality are irrelevant. Let
us call this hypothesis, which follows from the HO model, the Simple Games
hypothesis (SGH). Falsity of the HO model does not imply falsity of the SGH.
Even if the HO model is false because people fail to be rational in some situations,
the SGH might be true. Instead of viewing the ultimatum experiment as a test
of an already falsified very general hypothesis (the HO model), then, we may
consider it as a test of an non-falsified narrower hypothesis (the SGH).[17]

A test of the SGH presupposes that we can derive a prediction from it. If
the hypothesis is correct, participants' decisions in the ultimatum experiment
depend on preferences and expectations concerning the consequences of their
actions. If participants expect that their actions have no other material conse-
quences than the monetary payoffs explained in the instructions, they will try to
maximize their monetary returns. In the case of the responders, this means that
they will accept any positive offer. Proposers' offers depend on their expecta-
tions concerning responders' behavior. If proposers expect that responders will
accept any positive offer, they will never offer more than the minimal positive
amount.

Let us refer to this prediction—no offer above the minimal positive amount,
no rejection of positive offers—as the textbook prediction. It is well known
that the textbook prediction has been refuted in the original experiments and
hundreds of replications and variations. In these experiments, the most frequent
offer is an equal split, and the offer '80% for me, 20% for you' is rejected in
about fifty percent of all cases (see Camerer 2003, ch. 2, for a survey).

However, a straightforward falsification of the SGH by these experimental
facts would be possible only if the textbook prediction actually follows from
the SGH. This is, of course, not the case. The SGH must be supplemented by
several auxiliary hypotheses stating that, *for the given experimental design*, the
following conditions hold:[18]

---

[16] Bardsley at al. 2010, ch. 3, consider the ultimatum experiment as a test of game theory,
where hypotheses concerning participants' preferences occur as auxiliary hypotheses. Game
theory assumes rationality, which is also a part of the HO model, and equilibrium, that is,
coordination of players' expectations. For a test of game theory, only proposer behavior would
be relevant, because responders need not form expectations. In fact, the main point of interest
in the ultimatum game turned out to be responder behavior, implying that the ultimatum
game was mainly interpreted as a test of the HO model.

[17] On the relevance of this simple point for the methodology of economics, see Albert 1996.

[18] In order to derive the textbook prediction concerning responder behavior, no hypothesis

1. Participants have understood the instructions.

2. Participants believe that there will be no other material consequences of their decisions than the monetary consequences described in the instructions.

3. Proposers believe that responders have understood the instructions.

4. Proposers believe that responders are rational, egoistic, and motivated by material interests alone.

5. Proposers believe that responders believe that there are no other material consequences of their actions.

For each experimental design, we get different auxiliary hypotheses. These auxiliary hypotheses are nomological; they involve general claims for specific experimental designs.

Even though the textbook prediction fails in the ultimatum experiment, this leaves open the possibility that the SGH is correct and one of the auxiliary hypotheses is false. Of course, experiments are typically designed such that it is plausible that the corresponding auxiliary hypotheses hold.[19] The instructions are carefully explained to all participants until experimenters are confident that everybody has understood the rules. Moreover, in strategic contexts, one tries to ensure that participants are also confident that everybody has understood the instructions. Ideally, there are training rounds with questions testing the participants' understanding, and participants are informed that only those who pass the tests will participate in the experiment. Nevertheless, at some point, it must be assumed that the experimental design suffices to ensure that the participants' expectations satisfy the conditions described above.

The hypotheses concerning the effects of the experimental designs are a matter of debate and develop over time. Let me illustrate this for a special case.

In order to make sure that participants expect no further material consequences of their decisions in interactions after the experiments and outside the laboratory, experimenters try to convince participants that none of the other participants will ever learn about their decisions. This design feature is called anonymity (cf., e.g., Camerer 2003, 37–38). If anonymity is violated, the decisions in the experiment may be part of some larger game whose rules are unknown to the experimenter. Within such a larger game, any behavior observed in the laboratory could be consistent with the HO model.

Anonymity, however, is not enough to ensure that the game does not continue outside the laboratory. In most experimental designs, participants know that their decisions are observed by the experimenter. They may therefore play,

---

concerning proposer behavior is necessary, and the derivation involves only decision theory. Hence, for a falsification, we could ignore proposer behavior and the game-theoretic aspects. This does not mean that the falsification could have occurred in a decision experiment without a proposer. As we have learned from these experiments, it is responders' reaction to proposers' behavior that makes them reject money. Nevertheless, since proposer behavior belongs to the experimental facts that have to be explained, we consider all of the textbook prediction.

[19] On experimental design in economics, see, e.g., Camerer 2003, 34–42.

or believe to play, some larger game with the experimenter as an additional
player. In order to exclude this possibility, experiments must impose 'experi-
menter blindness' (EP). This is difficult to achieve. The experimenter must,
after all, get the data and make sure that participants stick to the rules and
receive the appropriate payoffs.[20]

In social psychology, the possible influence of the experimenter on partici-
pants' behavior is discussed under the heading of 'demand effects' (cf. Camerer
2003, 62): experimenters fear that participants try to 'help' them by doing what
they believe the experimenter would like them to do.

The HO model and, correspondingly, the SGH exclude the simple form of
demand effects in experiments with monetary incentives because the HO would
never sacrifice money just to help the experimenter. However, interaction outside
the laboratory could make it rational even for egoistic participants to do what
they think is expected from them. For instance, if participants are students
from the experimenter's classes, they might try to help the experimenter in
exchange for better grades. With participants that have no other dealings with
the experimenter, SGH-based demand effects are quite implausible.

Thus, EP is difficult to achieve, and demand effects are not very plausible on
the basis of the SGH. When we consider the ultimatum experiment as a test of
the SGH, implementing EP seems not very important. Once the SGH is con-
sidered as falsified, however, demand effects gain in plausibility: if participants
are not always selfish, they might sacrifice some money in order to do what they
believe is expected from them.

Strong effects of EP were, in fact, found in another experiment with mone-
tary incentives where EP is relatively easy to implement, namely, the dictator
experiment. In the dictator experiment, the proposer can distribute the money
as he pleases; the other player has no say. In this case, EP changed proposer
behavior in the direction predicted by the SGH (although the deviation from
the prediction was still considerable). This result was potentially troublesome
for the conclusions drawn from the ultimatum experiment. However, further ex-
perimentation showed that EP makes no difference in ultimatum experiments.

In order to conclude that the SGH is falsified by ultimatum experiments,
one should consider several such experiments, with different designs and, cor-
respondingly, involving different auxiliary assumptions. Moreover, results from
other experiments involving the same auxiliaries are also relevant. Taken to-
gether, the many relevant experiments not only provide a severe test and fal-
sification of the SGH. As already emphasized by Popper (1959), the so-called
experimental facts leading to a falsification are actually well-corroborated low-
level hypotheses. The importance of this point is obvious in the case of the
ultimatum experiment. Theories intended to replace the SGH or, more gener-
ally, the HO model must be consistent with these and many other experimental
facts; ideally, they should be able to explain them.

---

[20] On EP and its implementation in dictator and ultimatum games, see Hoffman et al.
1994; Bolton/Zwick 1995 and Camerer 2003, 62–63. EP was introduced in order to distinguish
between different non-HO explanations of the experimental facts. However, it is also potentially
relevant for tests of the HO model or the SGH.

This is just what theories of other-regarding preferences try to achieve. Fehr and Schmidt's (1999) theory of inequity aversion, for instance, tries to cover several experimental facts that seem, at first sight, hard to reconcile: in some experiments, participants behavior is quite consistent with the HO model, while in other experiments, behavior seems to be based on other-regarding preferences. Specifically, Fehr and Schmidt use data from ultimatum experiments to argue in favor of a hypothesis concerning the distribution of preferences that could explain the experimental facts.

3.3 Severe Tests and the Duhem-Quine Problem

The textbook prediction for the ultimatum experiment is derived by deduction from the SGH and several auxiliary hypotheses. When this prediction failed, several hypotheses could have been blamed since the prediction follows from the SGH and auxiliary hypotheses concerning the effect of the experimental design on the expectations of the participants. The failure of the textbook prediction implies that at least one of these premises is false, but it does not tell us which. In order to save the SGH from a falsification, one might argue, then, that one or more of the auxiliary hypotheses are false.

This is an illustration of the (weak) Duhem-Quine thesis, which claims that scientific tests usually involve several hypotheses, any of which could be to blame if a falsification occurs. The Duhem-Quine problem is the problem to decide which conclusions should be drawn from the falsification.[21]

Critical rationalism deals with the Duhem-Quine problem by requiring severe tests. A severe test of some target hypothesis presupposes that the auxiliary hypotheses themselves have been corroborated in severe tests. Only then can the falsification be blamed on the target hypothesis. If one of the auxiliaries is untested, any decision concerning the status of the target hypothesis must be delayed until the untested auxiliary has been tested. If one of the auxiliaries is falsified, the observation or experiment constitutes no test at all. The status of the auxiliary hypotheses is equally relevant for falsifications and corroborations.

Thus, the requirement of severe testing means that researchers who test some target hypothesis must use well-corroborated auxiliary hypotheses as an input. If they are successful, their output is again a well-corroborated hypothesis: either the target hypothesis is corroborated, or they produce well-corroborated low-level hypotheses, namely, the experimental facts that falsify the target. However, there is no guarantee of success. Implicit assumptions may be overlooked and turn out to be false. They may make logical errors. Experimental results may not be statistically significant. And so on.

The Duhem-Quine problem, then, means that, in testing hypotheses, corroboration is hereditary in the sense of *section 2*'s model.

---

[21] On the Duhem-Quine problem, see Gadenne 1998b. The strong Duhem-Quine thesis, which claims that any test involves all our beliefs, is exaggerated: quantum mechanics is not at stake in the ultimatum experiment. For a discussion of the Duhem-Quine problem in connection with experimental economics and, specifically, the ultimatum experiment, cf. also Bardsley at al. 2010, ch. 3.

3.4 The Logic of Scientific Discovery

A similar argument holds for the discovery or invention of new theories, which often proceeds by deductive arguments.[22] An example of such an argument is provided by the linearity assumption, which is used as a premise in many contexts. In the simplest case, it is conjectured that the connection between two measurable magnitudes is linear, and two observations are used to derive a new hypothesis, namely, a specific linear relation. A more complicated example involving statistics is provided by linear regressions. The history of science provides less trivial cases where more goes on than the specification of an unknown parameter.

The linearity assumption is a heuristic device for finding new hypotheses. The statement that, in some field of inquiry, starting with the linearity assumption is a good method for finding new hypotheses is a lower-level hypothesis of the same kind as the statement CR. It can be corroborated in a critical discussion.[23] Nevertheless, everybody knows that linearity may fail. The premises of such a heuristic argument may therefore lend some support to a new theory or hypothesis, although they cannot prove it to be correct.

Some developments in experimental economics illustrate this type of argument. Experiments like the ultimatum experiment lead to new experimental facts that have to be explained with the help of new hypotheses. Fehr and Schmidt (1999) use several arguments in favor of their new explanation. Specifically, they use ultimatum-game data to select a specific hypothesis from a set of hypotheses. The exact nature and acceptability of their argument is under debate (cf. Binmore/Shaked 2010; Fehr/Schmidt 2010). However, it is clear that Fehr and Schmidt try to support their new hypothesis by arguments where they invoke well-corroborated premises, for instance, the experimental facts established in many economic experiments, among them ultimatum experiments.

While the output of heuristic arguments based on well-corroborated experimental facts and widely accepted premises like the linearity assumption is not well-corroborated itself, it is more respectable than some theory that is not supported in this way. Researchers who would like to use or test some behavioral theory will focus rather on the theory of Fehr and Schmidt than some theory with less support. Again, this is a case of hereditary quality.

3.5 Further Duhem-Quine Worries

Consider the case of the original ultimatum experiment as a test of the SGH. Let us assume for the sake of simplicity that the auxiliary hypothesis that experimenter blindness is irrelevant in experiments with monetary incentives (henceforth, EPH) was the only untested auxiliary and that all the others auxiliaries were well-corroborated.

---

[22] See Musgrave 1999, ch. 15, for this point and examples from the history of science.

[23] Or it may be rejected in such a discussion. A good example for an assumption that had been widely accepted once but is now considered less convincing is the normality assumption in statistics.

Under these assumptions, it seems straightforward to argue that the experiment provided a severe test of the joint hypothesis SGH ∧ EPH. This joint hypothesis was falsified, then, but no clear-cut conclusion about the real target of the test, the SGH, could be drawn. When the EPH was falsified in the dictator experiment, the original experiment was implicitly shown to be no test of the SGH. The experiment implementing experimenter blindness in the ultimatum experiment, then, restored the falsification.

Corroborations, however, work differently, and this is worrisome. Let us shortly consider the hypothetical case of a corroboration of the joint hypothesis SGH ∧ EPH in the ultimatum experiment. Part 3 of the CA rule implies that all consequences of a corroborated hypothesis are also corroborated, implying that the SGH and the EPH would both have been corroborated.

A falsification of the EPH in the dictator experiment would, of course, have nullified the corroboration of the SGH. Still, the hypothetical corroboration is worrisome. If a successful prediction derived with the help of untested auxiliaries counts as a corroboration for the target hypothesis, researchers could resort to an "immunization strategy" (Hans Albert 1985[1968]) that protects the target against falsifications: make tests only with untested auxiliaries; in the case of successes, the target is corroborated; in the case of failures, the untested auxiliaries are considered as jointly falsified.

Given this problem, one might be tempted to rule out joint corroborations. However, it seems to me that scientific competition takes care of the problem. There are two related factors that reduce the attractiveness of the immunization strategy.

First, the more corroborations a target hypothesis earns over time, the more interesting would it be for a status-seeking researcher to shoot it down. This could be done by falsifying the untested auxiliaries involved in the corroborations. Each falsification would eliminate a corroboration.

Second, if researchers have internalized critical rationalism and the CA rule, they believe in corroborated but not in untested hypotheses. Thus, untested auxiliaries are a natural aim for severe tests because researchers are agnostic with respect to them while they believe in, or are at least more inclined to believe in, well-corroborated auxiliaries. Therefore, researchers would prefer to attack untested auxiliaries, and, for the same reason, to use corroborated auxiliaries.

Thus, not all explicit and implicit assumptions that are involved in deriving a prediction from a target hypothesis need to be corroborated. Scientific competition makes such a strict methodological rule unnecessary. This is a good thing. Researchers cannot completely avoid the risk involved in the use of plausible but untested auxiliaries, and they may always overlook problematic implicit auxiliaries. Moreover, from a personal and from a scientific perspective, it is a good strategy to publish an interesting idea or result before it is perfectly polished: one avoids to be scooped; critical attention is often good for one's career; and the resulting division of scientific labor leads to efficiency gains.

There exists another problem. In a test of an auxiliary hypotheses, further auxiliaries are needed according to the Duhem-Quine thesis. Does the require-

ment of severe testing together with the Duhem-Quine thesis lead to an infinite regress of testing?

Critical rationalists think that the Duhem-Quine problem can be solved, and it seems that Duhem would have agreed (see Gadenne 1998b). However, the exact solution remains a bit unclear; the severity requirement is only a local solution. Here is a proposal for a global solution: the theory of corroboration should take the network character of the set of hypotheses considered in a field into account.

The auxiliary hypotheses introduced in the context of the ultimatum experiment appear in several experiments. Moreover, they are often relevant for tests of competing hypotheses. For instance, it is often necessary to assume that certain procedures ensure that participants understand the instructions and are confident that the same is true of other participants. Hence, we find many hypotheses, some concerning human preferences, others concerning the effects of certain experimental designs, that are tested in varying combinations, leading to big clusters of experiments and experimental facts that have to be taken into account simultaneously in order to decide which hypotheses should be considered as corroborated.

Let me explain with the help of a formal example what is involved. Consider four hypotheses H_1, H_2, A_1, A_2. Let H_1 and H_2 be inconsistent, like different theories of human preferences. The other two hypotheses, A_1 and A_2, are consistent with each other and with H_1 and H_2; they might be concerned with the effects of certain experimental designs. Assume that H_i can be combined with A_j, $i, j = 1, 2$, yielding a low-level hypothesis P_{ij} that can be checked in an experiment. The form of these hypotheses is 'In an experiment with design X, participants do Y'. This situation is shown in figure 1.

Assume now that P_{11} and P_{12} have been falsified, while P_{21} and P_{22} have been corroborated, as indicated in fig. 1. It is logically possible that H_1 is true and H_2 is false; however, this would imply that A_1 and A_2 must be false. If H_2, A_1 and A_2 are false, the success of P_{21} and P_{22} is left unexplained. If one assumes, on the other hand, that H_1 is false and H_2, A_1 and A_2 are true, all observations are explained as far as this is possible within this network.

It seems to me that the CA rule should be strengthened accordingly: because the assumption that H_1 is true leaves too much unexplained, we should, in this situation, consider H_1 as falsified and H_2, A_1 and A_2 as corroborated. Rules for regulating the distribution of tentative assignments of truth values in a network of hypotheses, then, can solve the Duhem-Quine problem. The rule of maximizing the number of 'true'-assignments on the basis of given empirical facts seems to be a promising candidate for a completely general solution of the Duhem-Quine problem.[24]

---

[24] In such a network, it could be proved that the strong Duhem-Quine thesis is wrong. If one considers big networks where all currently entertained hypotheses are included, plausible network structures will not translate local assignments of truth values to have ripple effects throughout the network.
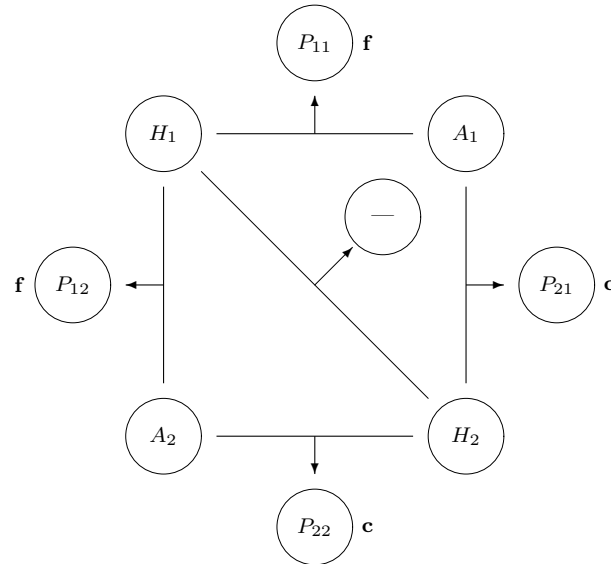
Figure 1: Network of nomological hypotheses. Circles stand for nomological hypotheses. Arrows stand for the relation of deductive consequence. The minus sign in a circle stands for a contradiction. The symbols **f** or **c** near a hypothesis indicates that the hypothesis has been falsified or corroborated.

4. Conclusion

In this paper, I have argued that critical rationalism is incentive compatible in the following sense. A researcher who proposes a new hypothesis should try to present deductive arguments in favor for the new hypothesis, arguments whose premises are, as far as possible, corroborated. Moreover, he should combine the new hypothesis with corroborated old hypotheses in order to explain known facts and to derive new testable implications. Ideally, he should also add empirical evidence supporting the new testable implications.

In other words, the researcher should put his new hypothesis into the network of hypotheses under discussion and make the most of the connections with those hypotheses that are corroborated. Successful explanations of known facts and, especially, new tests involving corroborated auxiliaries increase the chances of corroborating the new hypothesis. This, in turn, is a precondition for other researchers to use his hypothesis in similar deductive arguments.

In practical research, the premises of the deductive arguments are usually not stated completely, and there are different plausible sets of premises leading to the same conclusion. However, in subsequent critical discussions, implicit premises are discovered and criticized or defended, supporting the view that the original

argument was intended and is interpreted as a (possibly incomplete) deductive argument.

This analysis clarifies a crucial aspect of Hull's theory of science. Hull (1988, 310 and elsewhere) assumes that researchers draw upon the work of others in order to support their own work. Critical rationalism shows how support works and why everybody tries to support his work if he expects others to do it, that is, why support in this sense is part of an incentive compatible methodology. Moreover, this kind of support induces a division of scientific labor.

Specifically, the Duhem-Quine thesis, which has often been used to criticize critical rationalism by pointing to the fact that the conclusions from failed predictions are not obvious, turns out to work in favor of critical rationalism. The Duhem-Quine thesis says that we have to look at networks of hypotheses. A researcher who wants to corroborate his new hypothesis must use the corroborated hypotheses in the network as auxiliaries.

It may be doubted that any researcher wants his hypothesis to play the role of an auxiliary hypothesis in the work of other scientists. However, a hypothesis is an auxiliary hypothesis just from the perspective of a researcher who uses it in order to supplement his own new hypothesis for the purposes of explanation and prediction. Thus, depending on the context, even a prominent theory may be used as a source of auxiliary hypotheses. Indeed, given the progress of science, the optimal outcome from the perspective of a researcher is that his ideas and results become part of the accepted background knowledge, which is used by those concerned with new ideas and results as a reservoir of auxiliary hypotheses.

Bibliography

Albert, H. (1985[1968]), *Traktat über kritische Vernunft*, Tübingen, transl. as *Treatise on Critical Reason*, Princeton

— (2006), Die ökonomische Tradition und die Verfassung der Wissenschaft, in: *Perspektiven der Wirtschaftspolitik* 7, 113–131

— (2010), The Economic Tradition and the Constitution of Science, in: *Public Choice* 144, 401–411

Albert, M. (1996), Unrealistische Annahmen und empirische Prüfung, in: *Zeitschrift für Wirtschafts- und Sozialwissenschaften* 116, 451–486

— (2002), Der Kritische Rationalismus und die Verfassung der Wissenschaft, in: Böhm, J. M./H. Holweg/C. Hoock (eds.), *Karl Poppers kritischer Rationalismus heute*, Tübingen, 231–241

— (2003), Bayesian Rationality and Decision Making: A Critical Review, in: *Analyse & Kritik* 25, 101–117

— (2004), Methodologie und die Verfassung der Wissenschaft. Eine institutionalistische Perspektive, in: Held, M./G. Kubon-Gielke/R. Sturn (eds.), *Ökonomik des Wissens. Jahrbuch Normative und institutionelle Grundlagen der Ökonomik* 3, Marburg, 127–150

— (2006), *Product Quality in Scientific Competition*, Papers on Strategic Interaction 6–2006, Max Planck Institute of Economics, Jena

— (2007), Ökonomie der Methodologie. Eine institutionalistische Perspektive, in: *Universitätsreden* 70, Saarbrücken, 17–35, URL: http://wiwi.uni-giessen.de/dl/det/albert/12779

— (2008a), Introduction, in: Albert, M./D. Schmidtchen/S. Voigt (eds.), *Scientific Competition. Conferences on New Political Economy* 25, Tübingen, 1–9

— (2008b), Product Quality in a Simple OLG Model of Scientific Competition, in: *MAGKS Joint Discussion Paper Series in Economics* 04–2008, Giessen, URL: http://www.uni-marburg.de/fb02/makro/forschung/magkspapers/04-2008_albert .pdf

— (2009), Why Bayesian Rationality is Empty, Perfect Rationality Doesn't Exist, Ecological Rationality Is Too Simple, and Critical Rationality Does the Job, in: *Rationality, Markets and Morals* 0, URL: http://www.rmm-journal.de

— (2011), Methodology and Scientific Competition, forthcoming in *Episteme* 8

— (forthcoming), Von der vollkommenen zur kritischen Rationalität. Eine Kritik ökonomischer Rationalitätsauffassungen

Andersson, G. (1984), How to Accept Fallible Test Statements: Popper's Criticist Solution, in: Andersson, G. (ed.), *Rationality in Science and Politics*, Dordrecht, 47–68

— (1994), *Criticism and the History of Science. Kuhn's, Lakatos's and Feyerabend's Criticisms of Critical Rationalism*, Leiden

Bardsley, N./R. Cubitt/G. Loomes/P. Moffatt/C. Starmer/R. Sugden (2010), *Experimental Economics: Rethinking the Rules*, Princeton–Oxford

Binmore, K./A. Shaked (2010), Experimental Economics: Where Next?, in: *Journal of Economic Behavior and Organization* 73, 87–100

Bolton, G. E./R. Zwick (1995), Anonymity versus Punishment in Ultimatum Bargaining, in: *Games and Economic Behavior* 10, 95–121

Camerer, C. F. (2003), *Behavioral Game Theory*, New York–Princeton

Congleton, R. D. (1989), Efficient Status Seeking: Externalities, and the Evolution of Status Games, in: *Journal of Economic Behavior and Organization* 11, 175–190

Dasgupta, P./P. A. David (1994), Toward a New Economics of Science, in: *Research Policy* 23, 487–521

David, P. A. (1998), Common Agency Contracting and the Emergence of 'Open Science' Institutions, in: *American Economic Review* 88, 15–21

— (2004), Understanding the Emergence of 'Open Science' Institutions. Functionalist Economics in Historical Context, in: *Industrial and Corporate Change* 13, 571–589

Diamond, A. M. Jr. (2008), Economics of Science, in: Durlauf, S. N./L. E. Blume (eds.), *The New Palgrave Dictionary of Economics*, 2. ed., London

Fehr, E./K. Schmidt (1999), A Theory of Fairness, Competition, and Cooperation, in: *Quarterly Journal of Economics* 114, 817–868

— / — (2010), On Inequity Aversion. A Reply to Binmore and Shaked, in: *Journal of Economic Behavior and Organization* 73, 101–108

Gadenne, V. (1998a), Bewährung, Wahrheit und Akzeptanz von Theorien, in: Gadenne, V. (ed.), *Kritischer Rationalismus und Pragmatismus*, Amsterdam–Atlanta, 89–110

— (1998b), Spielarten des Duhem-Quine-Problems, in: *Logos* N.F. 5, 117–148

— (2005), Wozu normative Wissenschaftstheorie? Zur Notwendigkeit und Rechtfertigung von Rationalitätsprinzipien in der Wissenschaft, in: Gesang, B. (ed.), *Deskriptive oder normative Wissenschaftstheorie*, Frankfurt, 31–47

— (2006), Methodological Rules, Rationality, and Truth, in: Cheyne, C./J. Worrall (eds.), *Rationality and Reality. Conversations with Alan Musgrave*, Dordrecht, 97–107

Güth, W./R. Schmittberger/B. Schwarze (1982), An Experimental Analysis of Ultimatum Bargaining, in: *Journal of Economic Behavior and Organization* 3, 367–388

Hackl, F./M. Halla/G. J. Pruckner (2007), Volunteering and Income. The Fallacy of the Good Samaritan?, in: *Kyklos* 60, 77–104 (longer version: *Working Paper* 415, Department of Economics, Johannes Kepler Universität, Linz 2005)

Hoffman, E./K. McCabe,/K. Shachat/V. Smith (1994), Preferences, Property Rights, and Anonymity in Bargaining Games, in: *Games and Economic Behavior* 7, 346–380

House, C. L./E. Ozdenoren (2008), Durable Goods and Conformity, in: *RAND Journal of Economics* 39, 452–468

Hull, D. L. (1988), *Science as a Process*, Chicago–London

Jarvie, I. C. (2001), *The Republic of Science. The Emergence of Popper's Social View of Science 1935–1945*, Amsterdam–Atlanta

Marmot, M. (2004), *The Status Syndrome. How Social Standing Affects Our Health and Longevity*, New York

Mayer, Thomas (1993), *Truth versus Precision in Economics*, Aldershot

Mayo, D. G. (2010), Toward Progressive Critical Rationalism. Exchanges with Alan Musgrave, in: Mayo D. G./A. Spanos (2010), 113–124

— /A. Spanos (2010) (eds.), *Error and Inference*, Cambridge

Merton, R. K. (1973), *The Sociology of Science*, Chicago–London

Miller, D. (1994), *Critical Rationalism. A Restatement and Defence*, La Salle

Musgrave, A. (1993), *Commonsense, Science and Scepticism*, Cambridge

— (1999), *Essays on Realism and Rationalism*, Amsterdam–Atlanta

— (2010), Critical Rationalism, Explanation, and Severe Tests, in: Mayo D. G./A. Spanos (2010), 88–112

Popper, K. R. (1959), *The Logic of Scientific Discovery*, London

Roth, A. E. (1995), Introduction to Experimental Economics, in: Kagel, J. H./A. E. Roth (eds.), *Handbook of Experimental Economics*, Princeton, 3–109

Stephan, P. E. (1996), The Economics of Science, in: *Journal of Economic Literature* 34, 1199–1235

— (forthcoming), The Economics of Science, in: Hall, B. H./N. Rosenberg (eds.), *Handbook of Economics of Technical Change*

Strevens, M. (2011), Economic Approaches to Understanding Scientific Norms, forthcoming in *Episteme* 8

Vanberg, V. J. (2010), The 'Science-as-a-Market' Analogy: A Constitutional Economics Perspective, in: *Constitutional Political Economy* 21, 28–49