Rhetoric, Induction, and the Free Speech Dilemma*

Jesús P. Zamora Bonilla†‡

Scientists can choose different claims as interpretations of the results of their research. Scientific rhetoric is understood as the attempt to make those claims most beneficial for the scientists' interests. A rational choice, game-theoretic model is developed to analyze how this choice can be made and to assess it from a normative point of view. The main conclusion is that 'social' interests (pursuit of recognition) may conflict with 'cognitive' ones when no constraints are put on the choices of the authors of scientific papers, as in an 'ideal free speech situation'. Scientific institutions may help to solve this conflict. Lastly, some empirical predictions are offered that can inspire future social research of the refereeing process.

1. Introduction. 'Rhetoric' is the name we give to the strategic use of language. That language can be used in a strategic way means, at the very least, that we can choose what to say, and that saying different things will have different effects; on the other hand, our sayings may have different interpretations as well, and it is usually other people who decide how to interpret what we say. As a consequence, any series of sentences arranged in an intelligent conversation will be the result of the conversants' attempts to produce in the others some desired effects, and it becomes a suitable phenomenon to be studied from the point of view of game theory, i.e., that part of rational choice theory devoted to situations with two or

*Received September 2004; revised August 2005.

†To contact the author, please write to: Jesús P. Zamora Bonilla, Dpto. de Lógica, Historia y F. de la Ciencia, UNED, Humanidades Paseo de Senda del Rey 7, 28040 Madrid, Spain; e-mail: jpzb@fsof.uned.es.

‡Financial support from Fundación Urrutia Elejalde and from Spanish government's research projects PB98-0495-C08-01 and BFF2002-03353 is acknowledged. Previous versions of this paper were presented at the fourth congress of the Spanish Society of Logic and Philosophy of Science, and at the seventh congress of the Society for the Advancement of Economic Theory. Comments and corrections were received from Max Albert, Paco Álvarez, Christian List, Uskali Mäki, and Pascual Martínez Freire, as well as from two anonymous referees.

Philosophy of Science, 73 (April 2006) pp. 175–193. 0031-8248/2006/7302-0003$10.00

more interdependent agents. As it is well known, 'rhetoric' has been traditionally opposed to 'logic' (or even to 'science'), in the sense that, whereas the latter concept was assumed to refer to argumentation in the pursuit of *truth* alone, the former referred just to *persuasion*, independently of whether we really *believe* or not in the validity of our arguments or in the truth of our conclusions. Hence, under this traditional interpretation the realm of science would fall outside the scope of rhetoric, at least as long as scientific research is conducted in a 'proper' way. In the last twenty five years, however, a systematic attempt to study the process of scientific communication as an exercise in the strategic use of arguments has been carried out by a number of authors, leading to the conclusion that rhetoric (in this sophisticated sense) is not only present within science, but essentially belongs into its very core (e.g., Gross 1990; Pera 1994; Harris 1997). A deep disagreement still exists, nevertheless, about the consequences we must draw from this fact: for some philosophers, the idea that scientists' assertions are highly malleable may be dangerously close to the thesis that scientists systematically deceive, and so, a big part of contemporary research in epistemology might be interpreted as an inquiry into the (internal or external) *limits* to scientific rhetoric, i.e., into the cognitive or institutional mechanisms that may drive arguments in an epistemically 'sound' direction (e.g., Kitcher 1991, 1993; Goldman 1999). On the other hand, some sociologists, deliberately equating the notion of 'fact' with that of 'what scientists take to be a fact', have concluded that, since scientific *assertions* are the result of a 'rhetorical negotiation', so are the facts those assertions aim to represent (e.g., Knorr-Cetina 1981; Mulkay 1991).

In this paper I will try to show that a game theoretic analysis of the strategic use of language in scientific communication may provide a common ground for more 'conservative' and more 'radical' interpretations of scientific rhetoric. Like most formal models in social science, the one I will present here is undoubtedly a caricature of the real phenomena it tries to depict, but, as it is indeed the case with good caricatures, I hope it will reveal in a transparent way some essential aspects of those phenomena. Section 2 introduces the motivation of the model and its basic elements, which combine insights from the social study of science and from research on Bayesianism, game theory, and inductive logic. The formal core of the model is presented in Section 3, whereas Section 4 draws its normative consequences, the most important one being that, were the authors of scientific papers left to decide by themselves the interpretation of their discoveries, the quality of scientific knowledge would be non-optimal in a definite sense that is explained there. Finally, Section 5 connects again the results of the preceding analysis with the real world, presenting some empirical predictions derivable from the model, and

pointing to some scientific institutions as social mechanisms that help to solve this 'Free Speech Dilemma'.

2. Negotiating the Claims of a Scientific Paper. Whereas a traditional view depicted the ideal scientific report as simply an 'objective' and 'neutral' description of clearly repeatable experiments, we know, after a plenty of historical and sociological case studies, that the process leading to the acceptance of a scientific discovery is much more tortuous: there is always a certain degree of flexibility in how to present the outcomes of research, and the interpretations that are finally accepted are the result of a negotiation among several agents. One particular facet of scientific practice where 'flexibility' and 'negotiation' are clearly manifest is the process of peer review of scientific papers, and this has been one of the preferred fields for studying the 'construction of scientific facts'. Of course, most of the authors who have studied this process do not simply 'describe in an objective way' what they have found, but frame their 'discoveries' within a network of interpretive concepts that allow them to explain why the negotiations run in certain directions rather than in others. (These concepts are, of course, no less strategically chosen than the ones scientists employ.) One of the most interesting accounts was offered by Pinch (1985), whose basic ideas were developed by Myers (1990) in an impressive study of the refereeing process of scientific papers in biology.

According to the Pinch-Myers approach, the authors of the papers and their referees negotiate on the 'degree of externality' (Pinch) of the claims the papers contain, a concept that Myers explains as the 'distance' between those claims and the assertions that are already accepted within the discipline: the higher the externality level of a new claim, the larger the body of accepted claims it would be necessary to revise, were the new claim to be accepted. The fundamental point of disputation between authors and referees would consist in the conflict between the formers' attempt to maximize the 'externality' or 'novelty' of their own claims, and the latters' attempt to 'protect' the body of existing knowledge. An important resource in these disputations is the possibility to challenge the 'evidential significance' of every claim: on the one hand, previously accepted claims may provide arguments for doubting the empirical validity of the authors' assertions; on the other hand, authors (as well as referees) are usually able to offer different interpretations of their findings or ideas; stronger interpretations will have a higher level of externality but weaker empirical support, whereas weaker interpretations will have less 'externality' and more 'confirmation'. Of course, this explanatory framework is a gross simplification: to say the least, in practice more than two agents are always involved (editors, readers, co-authors, etc.), and disputations can refer to much more than the two dimensions selected by Pinch and Myers (e.g.,

the 'degree of externality' of one single claim may be different with respect to different fields, and because of different reasons that can be modeled as different variables). I think, nevertheless, that all relevant explanations, particularly in the social sciences, must simplify and idealize a lot, and the model presented in the next section will adopt the basic variables used by these sociologists, under an even more idealized framework.

Although Pinch and Myers provide excellent evidence showing that authors and referees behave in this way, their approach leaves several important things unexplained. This is more evident when we look at the author-referee interaction from a rational choice, game theoretic point of view. First, in the Pinch-Myers narrative, it seems that referees (or journal editors) act as discipline-guardians just in order to protect some kind of 'monopoly power' (the prevalence of the ideas and techniques they master), but this behavior makes plain sense only when authors do not belong to the core of the discipline and referees do: after all, the referees would like to be treated with more benevolence when they themselves submit a paper to a journal, particularly when the paper contains bold ideas. So, we need a clearer account of why scientists behave in such a different way when placed in the opposite poles of the 'peer' review process.

Second, just pointing to the goals of the agents does not provide an explanation of why they make the *specific* choices they make. For example, if authors may choose among a wide set of claims, why do they select one claim in particular, instead of others? Or, why do the referees end up accepting *some* papers after some revision, instead of rejecting them all, or accepting them only when their claims have been confirmed 'beyond any doubt'? And how do the choices of every agent *depend* on the choices of the others. (I.e., what is the 'equilibrium of the game', if there is one?)

Third, showing that the path leading to the acceptance of a scientific claim is circuitous and contentious does not entail *per se* that the claims accepted in the end are not the best ones that might have been adopted, or 'good enough' ones at least. Perhaps the flexibility and negotiation of interpretations is not an obstacle, but a *requirement* for getting scientific papers of a satisfactory quality. From my point of view, the sociological analysis of the refereeing process gives us too few hints to make a normative evaluation of the *epistemic quality* of this process's output.

In order to answer these questions, I suggest taking the Pinch-Myers' account not as an 'explanation' of the scientific rhetorical process, but just as an empirical description to be explained by an abstract model in which the actors are assumed to be rational players, simultaneously motivated by cognitive goals and social aims. With respect to the cognitive part of scientists' interests, my default assumption will be that it corresponds to a Bayesian 'epistemic utility function'; this assumption allows

one to interpret Pinch's concept of the 'evidential significance' of a theory or hypothesis t as just its conditional probability under the existing empirical evidence e ($p(t, e)$) and the 'degree of externality' concept as the 'content' of t, i.e., as the inverse of its prior probability ($1 - p(t)$). With respect to the social part of a scientist's utility function, I will assume that it basically refers to the desire of being recognized as the author of an 'important' discovery, i.e., as having proposed an hypothesis that becomes accepted by the scientific community, and that has the most substantial possible content.

3. The Free Speech Model. An economic model comprises two essential elements: the *preferences* of the agents whose behavior the model is about, and the *constraints* faced by the agents. The latter can be divided into 'natural' and 'social' constraints: natural constraints determine the actions agents have the physical or cognitive capacity to perform, whereas social constraints determine those actions the agents are allowed to perform (they are the 'institutional rules of the game'). In the following I will make a distinction between a *general* and a *specific* model. The former will contain assumptions only referring to the agents' preferences and natural constraints; the latter will also include an assumption about the rules according to which agents interact. In this section a very simple rule will be supposed: first, a scientist makes an assertion, and second, her colleagues accept or reject the claim contained in that assertion. Since no more constraints are put on scientists' decisions, I will call this model 'the Free Speech model', and my initial discussion will chiefly be focused on it, because it is more intuitive. Only in Sections 4.2 and 5 will I come back to the general model.

3.1. The Heuristic Function. Imagine a scientist ('the author') who has performed an experiment and is writing a paper reporting the results. According to the Pinch-Myers narrative, the author will have a set of options: she may interpret her results as very relevant, new, or even 'revolutionary', or just as more 'trivial' findings. As I said at the end of Section 2, this property of an hypothesis t will be identified with its *content*, i.e., $1 - p(t)$. On the other hand, the author is able to provide strong arguments showing that the results are empirically appropriate if they are interpreted as not very 'new', but has only weaker arguments for supporting more contentful interpretations. So, there is a negative correlation between the content of an interpretation and its degree of *empirical confirmation* ($p(t, e)$, e being the empirical evidence[1]). I will call the specific relation

1. The statement of the evidence, e, can perhaps be interpreted as the outcome of a previous 'negotiation'.

that tells us what is the maximum level of content the scientist may reach for a given level of empirical confirmation of her claims the '*heuristic function*'; so, the heuristic function tells us that $1 - p(t) = h(p(t, e))$. The relevant properties of h, assuming it is continuous and twice derivable, are the following (see Figure 1, at the end of this section):

(1.i) $h \geq 0$

(1.ii) $h' \leq 0$

(1.iii) $h'' \geq 0$

Inequality (1.i) means that, no matter what degree of confirmation is chosen, at least a null degree of content can be reached (e.g., by asserting a 'tautology'). (1.ii) expresses the inverse relation between confirmation and content; in general, this relation is usually caused by the fact that more contentful interpretations will be achieved by a stronger *generalization* of the results. Lastly, (1.iii) says that h is concave, i.e., equal increments in the confirmation level will demand higher and higher reductions in the content level, and vice versa (save when h equals 0 for some level of confirmation less than 1; in that case, h, h', and h'' will be 0 for all values of $p(t, e)$ beyond that level). In this section I will assume that the author is the only agent who knows the precise shape of h; this assumption will be relaxed in Section 5.

The heuristic function shows clearly that the flexibility of interpretation is not absolute, but *constrained*. The author would obviously *desire* that her results allowed her to announce a discovery with both $p(t)$ and $p(t, e)$ close to 1, but, unfortunately, she just cannot. *The connection between scientific claims and 'reality' is exhibited in the shape of the heuristic function*, whose values are always less than 1.[2] To say it with an economic metaphor, h is the 'epistemic constraint' of the author, as a consumer's income usually determines her budget constraint, or, more accurately, as a firm's output is constrained by its production function. The next question to answer is, of course, what point within h will be the author's optimum choice?

3.2. The Acceptance Function. As I said at the end of Section 2, our model assumes that authors want their results to become accepted as important 'discoveries' by the members of their disciplines (although in Section 6 some complications will be introduced). So, in order to deter-

2. Obviously, for $p(t, e) = 0$ (i.e., for fully disconfirmed interpretations of the data) the maximum content is 1, for the author can assert a logical contradiction. Nevertheless, for analytical convenience I will take $h(0)$ as being identical with the limit of h when $p(t, e)$ tends to 0.

TABLE 1. The Situation of 'Readers' When They Are Presented a 'Theory' Whose Degree of Confirmation Is $p(t, e)$ and Whose Content Is $1 - p(t)$.

	Claim Correct	Claim Incorrect
Accept claim	$1 - p(t)$	$-p(t)$
Reject claim	$-qp(t)$	k

mine what is the optimum interpretation the author may give to her results, we (and she) have to know how her colleagues are going to react if she chooses one interpretation or another. The situation of other scientists (the 'readers') when they are presented a 'theory' whose degree of confirmation is $p(t, e)$ and whose content is $1 - p(t)$ $[= h(p(t, e))]$, is shown in Table 1. For simplicity, it is supposed that all the members of the discipline (including the author) have the same preferences regarding whether to accept the theory or not, and even that they share the same subjective probabilities (although this assumption will be modified in Section 5); this assumption also allows us to take the numbers in this table as representing the utility function of an imaginary 'referee', someone who 'represents' the cognitive preferences of the discipline's members.

For readers, the utility of accepting or rejecting the theory depends on whether it turns out to be 'correct' or 'incorrect'. By a theory being 'correct' or 'incorrect' I do not mean that it is objectively true or false, since surely scientists will not be able to determine objective truth with absolute certainty. What is really important for scientists is that the theory 'works well enough', at least during the first years or decades after it is introduced. The figures in Table 1 are inspired, nevertheless, by the 'cognitive utility functions' employed in Bayesian approaches to scientific reasoning, particularly those of the first row, corresponding to the acceptance of t (see, e.g., Levi 1967): if the probability of t being right is $p(t, e)$, then the expected utility of accepting t equals $p(t, e) - p(t)$, a typical Bayesian measure of the cognitive value of a theory, which combines empirical support and information.[3] Table 1 differs from other Bayesian approaches in its analysis of 'rejection': on the one hand, if t is rejected and incorrect, the utility received by the readers will depend on the community's state of knowledge with respect to the problem t is intended to solve, and it is taken to be independent of $p(t)$ and $p(t, e)$; a high value of k represents a high confidence in finding some other solution. On the other hand, if t is rejected but right, the community will suffer a loss which depends on the informative content of t. Lastly, the factor q is a measure of scientists' attitudes towards risk: the higher the value of q, the smaller the values

3. A factor can be added such that the weights attached to truth and information are not the same, but I will ignore this complication here.

of $p(t)$ and $p(t, e)$ for which researchers will be ready to accept t (hence, the higher q is, the more 'risk loving' will scientists be). For simplicity, I will assume that q and k lie between 0 and 1.

From the point of view of the author of a scientific paper, the essential question is whether the theory she is going to propose will be accepted or rejected by her colleagues. The figures in Table 1 allow us to calculate for what values of content and confirmation will her theory be accepted ('EU' stands for 'expected utility'):

(2.i) $\mathrm{EU(accept)} = p(t, e)(1 - p(t)) + (1 - p(t, e))(-p(t))$
$= p(t, e) - p(t)$

(2.ii) $\mathrm{EU(reject)} = p(t, e)(-qp(t)) + (1 - p(t, e))k$

(2.iii) $\mathrm{EU(accept)} \geq \mathrm{EU(reject)}$ iff
$1 - p(t) \leq [(1 + k) - (1 + k + q)p(t, e)]/[1 - qp(t, e)]$

I will call $f(p(t, e)) = [(1 + k) - (1 + k + q)p(t, e)]/[1 - qp(t, e)]$ the 'acceptance function', for it tells us that the theory will be accepted if and only if the point of the heuristic function which is chosen $(x, h(x))$ is such that $h(x) \geq f(x)$. It is easy to check that f has the following properties (see Figure 1):

(3.i) $f' < 0$

(3.ii) $f'' < 0$

(3.iii) $f'(0) < -1$

(3.iv) $f(0) = 1 + k$

(3.v) $f((1 + k)/(1 + k + q)) = 0$

3.3. The Equilibrium. We are now ready to find the equilibrium of this simplified model. In our imaginary situation, the author announces first a point of her heuristic function (the other points of which are unknown to her colleagues), and later the other scientists decide whether to accept that 'theory' or to reject it. (For simplicity, I will assume that, if they are indifferent between accepting t or not, then they will accept it). Recall that, according to the sociological data mentioned above, scientists want to be recognized as discoverers of facts with the highest possible 'novelty', i.e., informative content. Hence, since the author wants her theory to be accepted in the first place, she has to choose a point of h for which $h \geq f$. If there is no such a point, she will just not present any theory. If f crosses h, then the author will choose that point on f at which h is maximal, in order to maximize the content of her theory. This situation is represented in Figure 1.

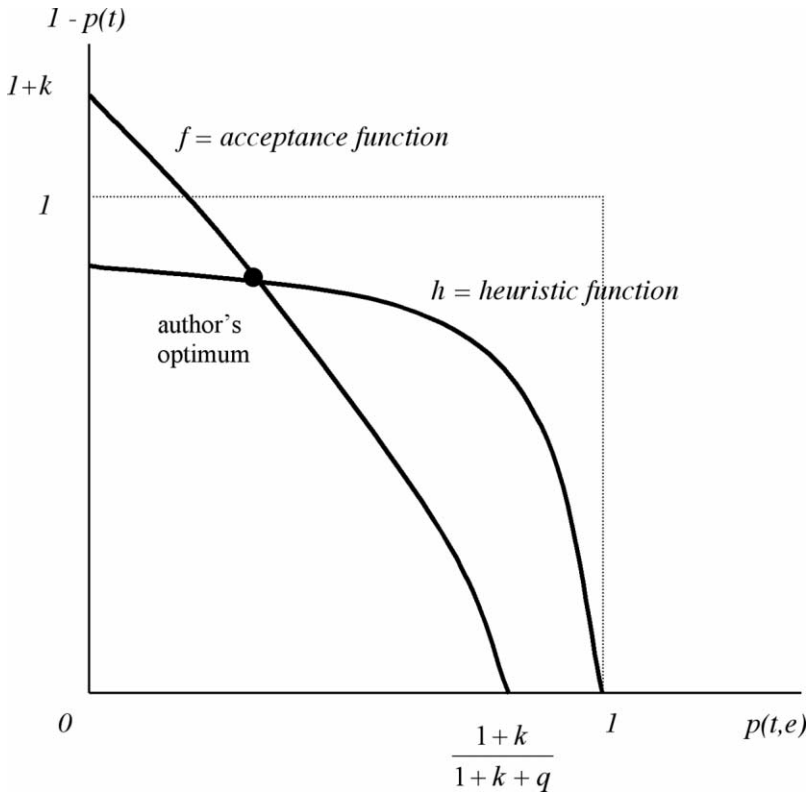

Figure 1. Equilibrium in the Free Speech model.

4. Normative Analysis.

4.1. The Readers' Curse. The Free Speech model describes the optimum choice of a claim by a 'recognition seeking' scientist to represent the results of her research, when this claim is going to be assessed by a group of 'Bayesian' colleagues. A crucial assumption is that 'recognition' or 'scientific merit' depends only on the claim's degree of content, if the claim is accepted. Some Bayesian epistemologists, or, in general, some rationalist philosophers, may doubt that recognition *should* have this form. The only answer I have is that, according to an impressive number of empirical studies, this seems to be how scientists are *actually* motivated, at least when they try to persuade their colleagues of the validity of their results. Nevertheless, I also find reasonable the hypothesis that scientists are basically driven by more neutral, cognitive goals, when *assessing* the

results offered by others, for this assumption has helped to explain many relevant features of scientific reasoning (see, e.g., Howson and Urbach 1989). So, I see no reason to deny that our assumptions about scientists' utility functions are empirically appropriate, in spite of the fact that some philosophers or sociologists may dislike one part of them or another. The *normative* analysis that will be made in this section is grounded, instead, on the principle that *de gustibus non est disputandum*: when studying a social situation, we are only allowed to give a normative assessment of it based on the preferences of the agents involved, not on *our own* preferences.

According to this principle, what we have to ask is, to put it bluntly: How good is the equilibrium point for the agents engaged in the game? Obviously, the equilibrium is optimum for the author, but what can be said about *readers*? On the one hand, it is clear that their decision of accepting the author's claim is also optimum for them, in the sense that they would gain nothing by changing their behavior, i.e., by rejecting the claim. (This fact is what makes the author's optimum a 'Nash equilibrium' of the game). But, on the other hand, perhaps readers would have preferred a different claim to begin with. We are going to see that this is *necessarily* so.

First, notice that (according to Table 1) the maximum utility that readers can get by *rejecting* a claim corresponds to the case where $p(t, e) = 0$ (i.e., when they are offered a fully disconfirmed hypothesis), which gives to them a utility level equal to k. So, the only points of the heuristic function that may be better than that one for readers are those for which the following is true:

(4.i) $\mathrm{EU}(\text{accept}) \geq k$ iff

(4.ii) $p(t, e) - p(t) \geq k$ iff

(4.iii) $1 - p(t) \geq (1 + k) - p(t, e) = i_k(p(t, e))$

The function i_k is the readers' indifference curve of utility k (if the claim is accepted). The properties of i_k are easy to deduce: it is a straight line of slope -1, crossing both axes at a distance $1 + k$ from the origin. In fact, all the indifference curves for the readers (from accepting the claim) are parallel to i_k, with utility growing as their distance to the origin increases. Hence we can derive the following:

(5.i) If $h < i_k$ for all values of $p(t, e)$, then the optimum point of the heuristic curve for the readers is $(0, h(0))$, and they would reject it (if offered by the author), getting a utility equal to k.

(5.ii) If $h > i_k$ for some values of $p(t, e)$, then the optimum point of the heuristic curve for the readers will be the one for which $h' =$

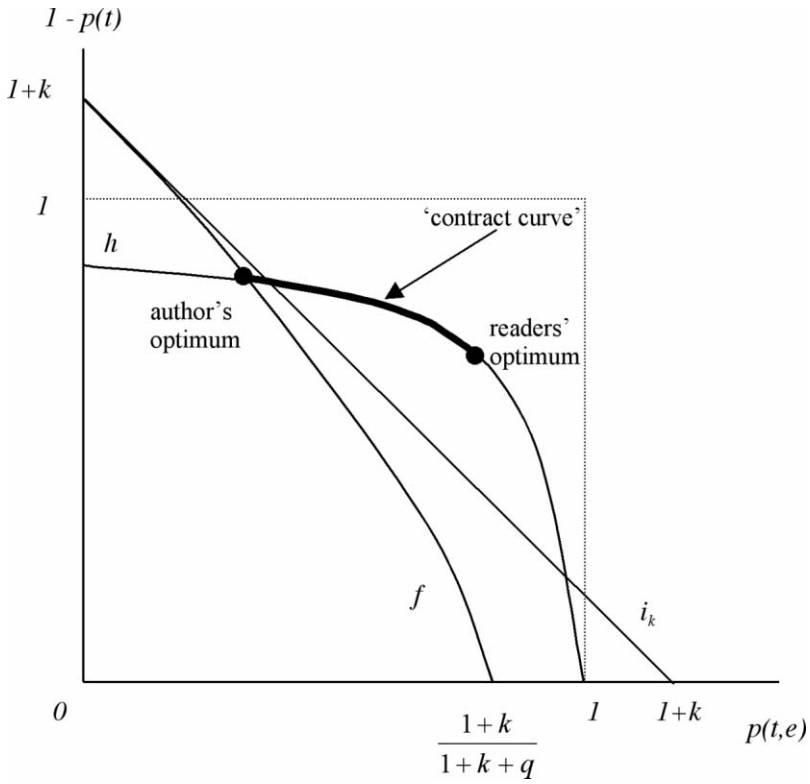

Figure 2. Author's and readers' optima in the normative analysis.

-1, and they would accept it (if offered), getting a utility higher than k.

(5.iii) The utility readers obtain at the author's optimum is less than k.

(5.i)–(5.ii) are straightforward. To prove (5.iii), it is only necessary to consider that, since $f' < -1$ and $f(0) = 1 + k$, then $f < i_k$ for all values of $p(t, e)$, and, being the author's optimum claim a point on f, it will give readers less utility than i_k. (There are two possible cases of indifference: first, when i_k is tangent to h at some point, that point also gives readers a utility of k; second, if $k = 0$ and $h' > -1$, their utility at $(1, h(1))$ is also k.) The case of (5.ii) is illustrated in Figure 2.

4.2. Efficiency: The Free Speech Dilemma. It comes as no surprise that the optimum for readers is not identical with the optimum for the author,

since, after all, the preferences of both types of agents are completely different. More unexpected, and more dramatic, is the conclusion that the claim that satisfies in the best way the preferences of the author leaves readers in a situation which is *worse* for them than if no claim where made at all (or, equivalently, than if a fully disconfirmed, and then easily rejectable, claim were made). I do not think this is actually the case in any real scientific field, not mainly because the idealizations and simplifications of Free Speech model make it 'unrealistic', but because actual scientific institutions *prevent* it from occurring (perhaps as an unintended consequence). I will discuss in the next section the role of institutions, as well as how a higher dose of realism would affect the consequences of our model; what I want to stress now is the fact that *the model describes a situation that seems to be ideal in a normative sense*, at least according to an opinion that is common amongst many critics of some real scientific institutions: authors are not constrained at all by 'the scientific establishment'; they know their colleagues' preferences perfectly; all published results become 'acceptable' for purely epistemic reasons; and authors receive all the glory they deserve. In a nutshell, the arrangement described in our model is as close as possible to a Habermassian 'ideal free speech situation'. And nevertheless, under those 'ideal' circumstances, *researchers would produce a corpus of scientific knowledge that, according to their own epistemic preferences, would be worse than having no knowledge at all!* This is what I propose to call 'the Free Speech Dilemma'.

Following our economic approach to this problem, the next thing we have to consider is whether the situation is *efficient*. In economics, an outcome is called 'inefficient' (or 'Pareto-inefficient') when there is at least one different situation in which at least one agent would have been better off, and no one worse off, than in the first situation; if this is not the case, the former outcome is 'Pareto-efficient'. So, our first question will be about what are the efficient points that readers and authors could reach (let us call the point on h for which the readers' utility is maximal from the decision of accepting the corresponding claim the "readers' acceptance optimum"; if h is always below i_k, then the readers' acceptance optimum is not a real optimum, because it gives utility less than k, whereas $(0, h(0))$ gives utility k):

> (6.i) If the author's optimum is to the left of the readers' acceptance optimum, then both points, as well as every point on h lying between them, are Pareto-efficient, and all other points are Pareto-inefficient.
>
> (6.ii) If the author's optimum is not to the left of the readers' acceptance optimum, then the author's optimum (it if exists) and all the points for which $p(t, e) = 0$ are Pareto-efficient, and all other points are Pareto-inefficient.

For the proof, note first that for any point below h and to the right of f, both the author and the readers would be made better off by moving upwards (save on the vertical axis, all of whose points are indifferent for readers and for the author); for points below h but to the left of f, moving to the left will make readers better off and leave the author indifferent. So, all points below h and for which $p(t, e) > 0$ are inefficient. Let us analyze the points on h. In the first place, if the author's optimum is to the left of the readers' acceptance optimum (as in Figure 2), then for all points to the right of the readers' acceptance optimum, this point is better both for the readers and for the author, whereas for those to the left of the author's optimum, the point $(0, h(0))$ is better for readers, and leaves the author indifferent. So, all points on h between the origin and the author's optimum, or to the right of the readers' acceptance optimum, are inefficient. With respect to those points between the author's optimum and the readers' acceptance optimum (if the latter exists), since all these points are to the right of f, the author's utility increases by moving to the left, and readers' utility grows by moving to the right, so it is impossible to make both better off simultaneously; so, all those points are Pareto-efficient.

In the second place, if the author's optimum is not to the left of the readers' acceptance optimum, note that the facts that the author's optimum is always below i_k, that $f' \leq -1$, and that $h'' < 0$, entail that in this case the full heuristic curve lies below i_k. Now, for all points to the right of the author's optimum, this claim is better both for the author and for the readers, whereas for all points to the left of the author's optimum, any point at the vertical axis is better for readers, and gives the same utility to the author. So, all points on h except the author's optimum and $(0, h(0))$ are inefficient. The author's optimum is Pareto-efficient because any other point will be worse for her, and all points for which $p(t, e) = 0$ are also Pareto-efficient because they give readers their maximum possible utility, k. This concludes the proof.

The most interesting case is the one considered in (6.i), which is again that of Figure 2. It entails that there is a kind of 'contract curve' between the author and the readers: the segment of h limited by their optima. *This segment is the space 'open for negotiation' between the author and the readers; i.e., this is the space for 'scientific rhetoric' to have its role*. Readers will try to persuade authors to make a claim as close as possible to the readers' optimum, whereas authors will try to stick as near to their own optimum as they can. In our Free Speech model, the equilibrium corresponds to the author's optimum because it is authors who make the choice of a claim and readers are assumed to ignore the shape of the heuristic function; but, under other institutional settings, the choice can be different. What the preceding analysis in this section shows is that, under any 'rea-

sonable' institution for negotiation, the equilibrium can be expected to lay on the 'contract curve'. Stated otherwise, this set of possible efficient equilibria is the 'solution' of the *general model* to which I referred at the beginning of Section 3, i.e., the model in which no assumptions about the social rules are made.

Regarding the general model, we have to extend the efficiency question to other conceivable ways to play the game. In order to do that, we must take into account that the *same* scientist can sometimes play the role of an author, and other times the role of a reader. Given the kind of interaction described in the Free Speech model, when placed as an 'author', a scientist cannot do better than playing her optimum, and when placed as a 'reader', she has no better choice than accepting the claim proposed by an 'author'. But suppose there were some way of *forcing* authors to propose the claim corresponding to the *readers'* optimum, or any other point of the 'contract curve'. We can then ask whether scientists, knowing that they will play both roles some of the times, might not prefer to be so *compelled* instead of having a free choice when acting as authors? In the 'ideal' circumstances inspiring the Free Speech model, all the members of a scientific community would have an even chance of being placed as 'the author' each time the game is played, and she would be 'heard' by all her colleagues. So, if there are n members in the group, on average each one will be an 'author' in a proportion of cases equal to $1/n$, and a 'reader' in a proportion equal to $(n-1)/n$. Let x be the utility attained by a reader. As we move to the right on the contract curve, x increases, and the utility of authors decreases. So, let us call $g(x)$ the utility an author gets when we choose a point of the contract curve that gives readers a utility equal to x. Hence, the expected utility of a researcher if that point were *always* chosen is

$$(7)\quad \mathrm{EU}(x) = x(n-1)/n + g(x)/n.$$

EU has a maximum when $g' = -(n-1)$, but, if $|g'| < n-1$ for all the values x may have on the contract curve, then EU will be maximized at the readers' optimum (i.e., when x is maximal). In conclusion, if the scientific community is big enough, or if g is not too steep (i.e., if the authors' utility does not decrease too intensely as she moves to the right on the contract curve), then the situation described in the Free Speech model will be inefficient, in the sense that the players themselves would prefer to play the game according to different rules; for example, they might prefer to be forced to choose the readers' optimum, instead of having the freedom to make their own choice when placed as 'authors'. (Of course, depending on the shape of g it is also possible that scientists systematically prefer the author's optimum, if the weight that recognition has in their utility function is much stronger than that of epistemic values,

but I do not think that it usually is, taking into account that scientists are readers much more often than they are authors). It is important to notice that this result does not contradict statement (7), for in that case we considered the efficiency of the possible outcomes of a *single* game played according to the rules of the Free Speech model, whereas we are now studying the efficiency of the *arrangement* according to which the game is played: any point of the contract curve is efficient in the sense that, once the author and the readers are playing, and the rules are given, there is no way of improving the utility of at least one agent without making others worse off; but, before knowing whether one is going to be an 'author' or a 'reader' in a particular case (i.e., considering the game 'under a veil of ignorance'), all scientists might prefer the readers' optimum, or some other point to the right of the author's optimum.

5. Back to the Real World. It is time now to confer more realism on our discussion. I will do so by two different, but related means. In the first place, I will try to derive some empirical predictions from the model, and I will also consider some possible changes in our assumptions about the cognitive capacities of the players, although I will leave for further research the formal analysis of the models deriving from these modifications. In the second place, I will refer to some alternative institutional arrangements of the interaction between authors and readers.

5.1. What Does the General Model Entail about Real Scientific Negotiations? Being a very abstract and simplified picture, we can hardly demand a full correspondence between our models and real scientific practice; in this way, these models are not different from most Bayesian theories, and I actually think their main virtue is to show how the existence of 'rhetorical negotiations' can be accommodated within a Bayesian approach, or, in general, within a rational choice approach. In particular, one basic handicap for the empirical testing of our models is that the nature of the existing data about refereeing in scientific journals (the part of the publication process more closely related to 'rhetorical negotiation') does not obviously fit the concepts with which Bayesian models are built up. Nevertheless, I think at least some 'predictions' can be made, that might be tested through the future accumulation of more sophisticated data. I will offer only 'predictions' derivable from the *general model*, i.e., only based on the assumption that some efficient equilibrium (a point of the 'contract curve') is reached through the negotiation between authors and readers.

To begin with, I will derive an empirical claim that is patently false, and that calls for the removing of some of the model's idealizations:

(8) All the claims actually presented by authors will be accepted by readers.

It is clear that many papers are not accepted for publication, and when they are, their claims are later rejected (or simply ignored) by the author's colleagues. Our model has this unrealistic consequence because it assumes for the author a perfect knowledge of the readers' epistemic preferences, and identical preferences of all readers as well. If these assumptions are dropped, then each possible claim would be associated with some *probability* of getting the claim accepted, or, even more realistically, with some *expected number* of colleagues accepting the claim. Most choices under those circumstances would lead to some positive level of rejection.

The following 'predictions' are more plausible ('a' stands for the author's optimum claim):

(9) If there is a possible claim t on h such that $p(t, e) > 0$ and such that readers prefer to accept t rather than a, then $p(t) > p(a)$ and $p(t, e) > p(a, e)$.

The antecedent of (9) is equivalent to the assertion that the readers' acceptance optimum is to the right of a (as in Figure 2). What this proposition asserts is that authors will tend to propose claims more contentful and less empirically confirmed than the claims readers would prefer. Of course, our model has been designed to have this consequence, and so it is not a real 'prediction', but rather an 'explanation' of the facts described in Section 2. But it could be taken as a 'prediction' in the sense that it is not clear *a priori* that (9) must be the case: my assertion would be falsified by finding that, under certain circumstances, readers tend to prefer claims that are less 'modest' than those originally presented by the authors. Sociologists of science are invited to look for such cases.[4]

The following prediction, instead, is less compatible with the sociological theses studied in Section 2:

(10) Under the same assumption, if $h(1) = 0$, then readers will not prefer a claim that is maximally confirmed.

The new condition in the antecedent of (10) is that, for claims with a

4. The editor of *Philosophy of Science* has commented that the referees of the journal often believe that authors do not emphasize enough the novelty of their claims, and even in these cases the referees may recommend the publication of the paper. I accept that this can be taken as a point against my 'prediction', and that some explanation must be offered of why authors are sometimes too modest. For example, uncertainty about the referee's cognitive utility function may lead authors to choose a 'play safe' strategy; I guess that this uncertainty would be higher in areas like philosophy than in natural science.

positive degree of information, the heuristic curve only allows any evidence to confer them a degree of confirmation less than 1 (what seems to be a reasonable assumption). In that case, what (10) asserts is that readers will not demand a high verification level 'at any cost', but will prefer some claim that has a positive level of content. This result may explain why referees, although depicted as 'orthodoxy guardians' by sociological accounts, nevertheless accept claims with a positive degree of 'externality', in Pinch's terms.

The following two predictions are more novel.

> (11) Under same assumption as in (9), as k grows, the size of the contract curve diminishes from the left.

Recall that k was a measure of the utility readers expected to derive from other possible solutions to the problem the author was trying to solve. What (11) asserts is, then, that the more optimistic scientists are about the prospects of finding a solution to that problem, the less willing they will be to accept a claim that is to the left of their own optimum on the heuristic curve. Stated in Kuhnian terms, when there is a great confidence in the efficacy of the 'paradigm', the members of a scientific community will demand that the claims presented by their colleagues are very close to the readers' optimum, and 'negotiations' will be very tough on the part of referees; in turn, when the confidence is not so strong, referees will often accept more innovative claims. A different reason why the value of k may be higher is that there exists a strong *competition* amongst researchers. In this case, what the prediction asserts is that, the higher the concurrence for finding the solution of a scientific problem, the more intensely will referees insist in that an acceptable claim must be close to the readers' optimum.

Finally,

> (12) If a is preferred by readers to all other points on h better confirmed than a, then referees will try to reject the author's claim.

Obviously, the antecedent of (12) is the opposite of that of the other predictions: it says that the readers' acceptance optimum is to the left of a, or it is identical to a. As we saw in the proof of (6.ii), this condition entails that h is below i_k, and then the readers' global optimum is $(0, h(0))$ (i.e., no claim is made). What our general model predicts is, then, that, if referees *suggest* that the author's paper would improve by making some more contentful, but less confirmed, claim, then they or the editors will finally decide to reject the paper. I think that the truth of this prediction is less clear from the knowledge we have about the refereeing process, either by means of social studies, or simply as participants, and its testing will hence be of the highest interest (but see note 4).

The sociological studies inspired by the testing of these predictions will also help to reformulate and improve the model presented here in a number of ways. For example, some observable magnitudes should be defined to take the place of our Bayesian, unobservable variables $p(t,\ e)$ and $1 - p(t)$, as well as k and q. The heuristic function could also be defined not as a continuous curve, but on a finite set of possible claims, those actually conceivable by the members of the scientific community. Likewise, we could explore the possibility that authors and readers perceive a different heuristic function (e.g., authors may probably suffer from 'confirmation bias', thinking that the values of h are systematically higher than those readers believe it has). Lastly, more realistic and complicated assumptions about scientists preferences could be made with the help of empirical studies. These modifications of our general model should not be taken as *ad hoc*, but as steps in a strategy of 'de-idealization', to use Nowak's concept, or as an element of the 'positive heuristic' of a research programme, to say it with Lakatos (e.g., Brzesinky and Nowak 1992; Lakatos 1977).

5.2. Alternative Institutional Settings. My final comment will be about how our general model can be specified in ways that resemble more closely the actual institutions of scientific publication. Recall that the Free Speech model described a setting that was the most beneficial one for those scientists playing the role of authors. I do not think this setting exists in any area of the realm of science, but I fear that many are hoping that electronic publication will drive scientific communication towards something like an 'ideal free speech situation'. The model described in Sections 3 and 4 shows, however, that if this hope were *really* fulfilled, it would be destructive for the epistemic value of scientific knowledge, at least if scientists were driven by the mere pursuit of recognition when acting as authors. This point need not be taken as a criticism of electronic publishing, but of the idea that the absence of institutional constraints on scientific communication is necessarily a good thing. My point is simply that, as long as the preferences of a scientist as an author and as a reader are not identical, establishing a social mechanism that systematically favors the author's interest will tend to worsen the quality of scientific knowledge.

We can think, hence, of institutions like refereed journals as ways of 'forcing' authors to make claims that are closer to the optimal ones from the point of view of an average member of the scientific discipline. Obviously, these institutions may suffer from other problems, and will be more difficult to analyze; for example, in this case we have to distinguish the role of 'readers' from those of 'editors' and 'referees', and new conflicts may arise between their respective interests. Research on the formal modeling of these complicated interactions is surely needed. But the most

beneficial outcome of this type of formal analysis is that it helps us devise new institutions to regulate scientific communication. A simple suggestion can be derived from proposition (9): scientific competition restricts the set of efficient equilibria to those points closest to the readers' optimum, and hence improves the quality (from the readers' viewpoint) of the claims presented by authors. Those institutions that foster competition (technically understood as the likelihood that 'rival' researchers solve the problem you are trying to solve before you) will have a beneficial effect on the quality of scientific knowledge. Nevertheless, the analysis of all these institutions has to take into account not only their epistemic efficiency, but also all other kinds of costs and benefits involved in the production and communication of knowledge, particularly those costs and benefits experienced by the common citizens.

REFERENCES

Brzezinski, J., and L. Nowak (1992), *Idealization III: Approximation and Truth*. Poznan Studies in the Philosophy of Science, 25. Amsterdam: Rodopi.

Goldman, A. (1999), *Knowledge in a Social World*. Oxford: Oxford University Press.

Gross, A. G. (1990), *The Rhetoric of Science*. Cambridge, MA: Harvard University Press.

Harris, R. A., ed. (1997), *Landmark Essays on Rhetoric of Science: Case Studies*. Mahaw, NJ: Lawrence Erlbaum.

Howson, C., and P. Urbach (1989), *Scientific Reasoning: The Bayesian Approach*. La Salle, IL: Open Court.

Kitcher, P. (1991), "Persuasion", in M. Pera and W. R. Shea (eds.), *Persuading Science: The Art of Scientific Rhetoric*. Canton, MA: Science History Publications, 3–27.

——— (1993), *The Advancement of Science*. Oxford: Oxford University Press.

Knorr-Cetina, K. (1981), *The Manufacture of Knowledge: An Essay on the Constructivist and Contextual Nature of Science*. Oxford: Pergamon.

Lakatos, I. (1977), "Falsification and the Methodology of Scientific Research Programmes", in *The Methodology of Scientific Research Programmes*. Cambridge: Cambridge University Press, 8–101.

Levi, I. (1967), *Gambling with Truth*. New York: A. Knopf.

Mulkay, M. J. (1991), *Sociology of Science: A Sociological Pilgrimage*. Milton Keynes, UK: Open University Press.

Myers, G. (1990), *Writing Biology*. Madison: University of Wisconsin Press.

Pera, M. (1994), *The Discourses of Science*. Chicago: University of Chicago Press.

Pinch, T. (1985), "Towards an Analysis of Scientific Observation: The Externality and Evidential Significance of Observational Reports in Physics", *Social Studies of Science* 15: 3–36.