Draft as of March 2015
The final version of this paper was published in *The Epistemic Life of Groups: Essays in the Epistemology of Collectives*, ed. by Michael S. Brady and Miranda Fricker, Oxford: Oxford University Press, pp. 218-233.
https://doi.org/10.1093/acprof:oso/9780198759645.003.0012

Torsten Wilholt
Institute of Philosophy, Leibniz Universität Hannover
torsten.wilholt@philos.uni-hannover.de

Collaborative Research, Scientific Communities, and the Social Diffusion of Trustworthiness

Abstract

The main thesis of this paper is that when we trust the results of scientific research, that trust is inevitably directed at least in part at collective bodies rather than at single researchers, and that accordingly, reasonable assessments of epistemic trustworthiness in science must attend to these collective bodies. In order to support this claim, I start by invoking the collaborative nature of most of today's scientific research. I argue that the trustworthiness of a collaborative research group does not supervene on the trustworthiness of its individual members and point out some specific problems for the assessment of epistemic trustworthiness that arise from the specific nature of today's collaborative research. Next, I argue that the social diffusion of trustworthiness goes even further; we always also need an assessment of the trustworthiness of the respective research community as a whole. Communities, I claim, play an essential role in the epistemic quality management of science. To see why this role is indispensible, we have to appreciate the full complexity of determining what is desirable in a method of inquiry. The relevant features of a method include three different dimensions: the reliability of positive results, the reliability of negative results, and the method's power. Every methodological choice involves a trade-off between these three dimensions. The right balance between them (the "distribution of inductive risks", or DIR) depends on value judgments about the costs of false results and the benefits of correct ones. Conventional methodological standards of research communities impose constraints on admissible DIR and thereby harmonize the implicit value judgments. Trusting that the research community has set the limitations on DIR in a suitable way is thus always part of placing our trust in a scientific result.

1. Introduction

When we invest epistemic trust in results of scientific research, our trust is often directed at a collective body rather than at a single researcher. As an empirical observation, this should be quite uncontroversial. In this paper, I will argue that this empirical fact is not just due to

convenience or habit: With regard to scientific information, every reasonable assessment of trustworthiness *must* attend to collective bodies rather than only to single researchers.[1]

Claims very close to this one have been defended with good reasons before, in particular by Deborah Tollefsen (2007),[2] but I will attempt a new approach to the problem. The novelty lies in the precise nature of the arguments that I will use to support the claim and, perhaps more importantly, the kinds of collective bodies that I have in mind.

In a first step I will appeal to the collaborative nature of most of today's scientific research. The trustworthiness of a collaborative research group does not reduce to the trustworthiness of its individual members. Again, this is very much in line with observations offered by other social epistemologists. But I will try to point out a few aspects of the nature of collaborative research that I think deserve special attention in this context.

Furthermore, I am going to argue that the social diffusion[3] of trustworthiness goes even further and extends beyond the level of collaborative groups: Our epistemic trust is, and needs to be, also directed at entire research communities defined by shared methodological standards. As long as we regard methodological standards only as means of codifying and putting on record the procedures that are most suitable for arriving at reliable results, this sort of trust in research communities might appear to be merely a practical contingency. After all, collaborative groups or even individual researchers might be regarded as "in principle" individually responsible for finding out by themselves which methods are the most reliable ones. However, I am going to argue that important methodological standards are in many cases solutions to problems of coordination rather than optimization. They are thus conventional and irreducibly social in character.

I am also going to argue that the conventional standards by which research communities are defined are crucially important for those features of the research that determine its trustworthiness. Therefore, any reasonable assessment of trustworthiness must involve a focus on the research community.

2. The collaborative nature of scientific research

Collaborative research is on the rise. An indicator of this is the increasing number of authors per scientific paper. Co-authorship is a symptom of the collaborative nature of research that is particularly important in our context, because a published paper is arguably still the most important medium through which a collaborative research group functions as a provider of

[1] Trustworthiness is commonly attributed to either a piece of information or a source of information. I take it that the latter is the more basic kind of trustworthiness-attribution, because usually the attribution of trustworthiness to a piece of information at least partially derives from an attribution of trustworthiness to its source. In any case, in this paper, when I use the word trustworthiness, I mean the trustworthiness of a source.

[2] While she has framed her arguments in terms of the question of what constitutes the testimony of a group, the above claim about trustworthiness could be regarded as a corollary of her thesis.

[3] The phrase "social diffusion" and its use to indicate non-supervenience on the properties of individuals is shamelessly lifted from Goldberg 2006.

information (und thus becomes the potential recipient of epistemic trust). Extreme illustrations of the phenomenon of co-authorship are easy to find. An article on one of the particle detectors of the Large Hadron Collider lists 2,926 authors (The ATLAS Collaboration et al. 2008). A paper on a large clinical trial comparing different thrombolytic drugs for the treatment of heart attacks cites the name of the collaborative research group as author; 972 investigators are named in an appendix to the paper (The GUSTO Investigators 1993). Multiple authorship is a general trend in many disciplines: In the US, the average number of authors per paper in the medical sciences increased from 3.7 to 6.0 in the twenty years leading up to 2010; in physics the number more than doubled in the same time (from 4.5 to 10.1), in astronomy it more than quadrupled (from 3.1 to 13.8).[4]

When scientific results are generated through collaborative research, the trustworthiness of the group cannot be reduced to the trustworthiness of its members. To see this, first note that the trustworthiness of a group is certainly not simply a function of the average trustworthiness of its members, or of the trustworthiness of its least or most trustworthy member. Familiar arguments from social epistemology (Tollefsen 2007; cf. Surowiecki 2005) make this clear: The competence of a group can exceed the competence of even its most competent members, as can be illustrated by Condorcet's jury theorem and other considerations pertaining to the wisdom of crowds. Similarly, no single member of a collaborative research group has all the information that constitutes the basis for the competence of the group as a whole.[5]

Taken by themselves, these points might seem to leave open the possibility that the trustworthiness of a group may as a matter of fact somehow supervene on the trustworthiness of its members (if not by way of any simple and straightforward functional dependence). This might in fact be the case in a group that somehow forms its group testimony by a method of aggregation from the views of its members. But in many interesting cases, and typically in the case of a scientific research group, that is not so. In our context, an additional reason why the trustworthiness of the group does not reduce to the trustworthiness of its members should receive particular emphasis: A group's trustworthiness crucially depends on its social organization. More specifically, this concerns facts about how information is distributed among its members and processed within the group (including practices of deliberation among its members) and how epistemic trust *within* the group is enabled and maintained.[6]

---

[4] Numbers from National Science Board 2012, p. 5-36, figure 5-24; see the underlying data at http://www.nsf.gov/statistics/seind12/c5/fig05-24.xls. (Accessed 25 March 2015).

[5] The last point is particularly important for groups that bring together expertise from different areas, sometimes different disciplines. The combined competence of a research group is often much greater than the competence of an individual scientist could ever be, as philosophers of science have often observed (Hardwig 1985, Thagard 1997, Wray 2002, 2006).

[6] Among these features of a group I would also count the degree to which the individual members of the group are approaching the group's epistemic objectives from slightly different angles. As philosophers of science have emphasized (esp. Longino 1990, 2002), the mechanism of mutual criticism is highly relevant for scientific knowledge generation. The diversity (or, as the case may be, homogeneity) of a group with regard to background assumptions, methodological preferences and the like is thus likely to play a role for its overall trustworthiness as a provider of robust scientific results.

These observations have important consequences for the practical problem of assessing the trustworthiness of a group. With regard to the possibility of establishing and maintaining trust in the kind of groups that we encounter in the sciences, two features of present day collaborative research deserve particular attention.[7] First of all, collaborative experiments or studies are often multi-site, i.e. they are spatially dispersed and involve many institutions. This limits our ability to assess the trustworthiness of the group with the aid of institutional indicators (such as reputation and track record of institutions). The spatial dispersion, in connection with the size of the group, may also present difficulties for discerning and understanding the social organization of the group. The social organization of a large randomized clinical trial, for example, may often not be transparent even to most of the participating researchers themselves. Secondly, many groups exist only for the duration of a single experiment or study (if by "group" we mean the collective body that produces a given piece of scientific information). This makes it impossible to assess the trustworthiness of such groups with the aid of their track record.

It would seem that both these points refer us back to an assessment of the trustworthiness of a group's individual members. After all, with individual scientists we *can* use cues like their institutional affiliations and other indicators of their standing within the community as indirect evidence for their trustworthiness, and we can try to assess their track record. But, of course, the arguments we have considered just before remain unrefuted: The trustworthiness of a group cannot be derived from the trustworthiness of its members.

In its present form, the collaborative character of scientific research therefore presents us with considerable difficulties when it comes to assessing the trustworthiness of scientific information. Estimating the trustworthiness of a collaborative group as a source of information can be very taxing even for scientific peers, and will usually overtax the abilities and capacities of extra-scientific users of scientific information.

3. The level of research communities

The previous section has revealed problems for a case-by-case assessment of the trustworthiness of particular sources of scientific information. This already seems to provide a good pragmatic explanation for the phenomenon that epistemic trust in scientific results is often grounded in an attitude of trust towards science as a whole, or towards particular research communities.

However, as noted at the outset, I want to argue for a stronger claim than can be supported by such pragmatic considerations: A reasonable assessment of the trustworthiness of a source of scientific information *must* draw on an assessment of the trustworthiness of the pertaining research community as a whole. My conviction about this stronger claim rests on additional reasons that are by and large independent from the considerations pertaining to the collaborative nature of scientific research, as will become clear in what follows.

[7] Both these features and their epistemic significance have been stressed by Rebecca Kukla and Bryce Huebner in their paper "Making an Author: Epistemic Accountability and Distributed Responsibility", presented in 2011 at both the SPSP conference in Exeter and at the EPSA in Athens; cf. Kukla 2012 and Winsberg et al. 2014.

Research communities in the sense of this paper are constituted by shared methodological standards. By methodological standards, I mean a wide range of guidelines and instructions for scientific practice that come in differing degrees of explicitness, generality and binding force and can govern all kinds of steps in the research process, from experimental design to data analysis to the dissemination of information. The intended scope of my use of the concept can best be illustrated by way of example. One methodological standard that is widely shared by several research communities is to regard 0.05 as the highest reasonable significance level in significance tests. Another standard, far more restricted in the scope of its application, is the following: In animal experiments in toxicology, you should always have a positive control group in order to test whether your experimental setup is sensitive to the kind of intervention you are investigating (cf. NTP 2001: vii). The biomedical research community has by and large adopted the following rule, which should be regarded as a standard concerning the dissemination of research information: Before you start a clinical trial, you should register it with a public repository, including information on endpoints and study design (cf. ICMJE 2010: 11, WMA 2008: 3). As a fourth and final example, consider the following *proposed* standard, which was recommended in a discussion paper resulting from a workshop held by the German Association for Plant Biotechnology (Broer and Schiemann 2009): In biosafety research on possible interactions between genetically modified organisms and their environment, every experiment should start from a scientifically-based causal hypothesis.[8] The examples illustrate how standards range from the highly specific to the near universal, as well as from the uncontested common ground to the controversial suggestion.

Obviously, research communities shape the trustworthiness of scientific information by means of the standards they implement. In that sense, our trust in a piece of scientific research may at least be in part directed at the respective research community and their collective competence in setting the appropriate standards. But so far as this consideration goes, this might still be considered to be just a contingent matter. It might seem that that conclusion could be arrived at in the following way: In principle, there exists *the* most trustworthy scientific method for any given problem. The trustworthiness of any particular piece of information depends on whether it was arrived at using the most trustworthy methods. The role of research communities and their institutional organs (like working groups, peer review panels and the like) is only to formulate guidelines that help individual researchers to identify the most trustworthy procedures. The collective efforts of research communities thus help to avoid the unnecessary duplication of individual cognitive efforts for identifying the most trustworthy methods, but their contribution is in no way essential or indispensable. Again, the community level would bear its trust-enabling role only for pragmatic reasons.

I think that the first step of this line of reasoning already contains a mistake: Typically, there is no such thing as the objectively most trustworthy method for a given problem. I therefore

[8] The discussion paper "Biologische Sicherheitsforschung an gentechnisch veränderten Pflanzen", written by Inge Broer and Joachim Schiemann in October 2009, was circulated widely in the plant biotechnology community and publicized via the website of the German federal ministry of education and research; the proposed standards continue to be discussed, cf. Broer 2012.

hold that in many important cases, methodological standards do not simply codify a method that can be regarded as the objectively most trustworthy one. Instead, they are conventional in character and thus irreducibly social.

My take on standards thus runs counter to a widespread view on methodology and epistemic quality management in general. According to this view, we can pass a purely epistemic judgment on methods or types of inquiry by telling how good they are at getting at the truth, and this decisive quality of a method is measured by its reliability. The most trustworthy method for tackling a given problem would therefore appear to be the most reliable one. In the following section, I will argue that reliability cannot play the role it is assigned by this view. Methodological choices are underdetermined by the aim of arriving at reliable results.

4. Is there such a thing as the reliability of a method?

My argument starts from a very down-to-earth reflection that is familiar to every methodologist. In every empirical investigation, there are at least two different kinds of way in which the investigation can go wrong: It can lead to a result which indicates that *p*, while *p* does in fact not obtain (call this a false positive), or it can fail to lead to such a result, while *p* is in fact true (false negative). If you are conducting an open investigation into the question of whether *p*, you do not yet know whether the true result is positive or negative. But the reliability of positive results and the reliability of negative results of one and the same method of inquiry can differ widely. For example, if a rapid HIV test was performed on a *random group* of German citizens, negative results would have a reliability of over almost 100 percent, while positive results would have a reliability of well under 50 percent (that is, almost 100 percent of those who would receive a negative result would in fact be free of the HIV retrovirus, while the proportion of HIV carriers among those who would receive a negative result would be under 50 percent). So if you wanted to judge a method by its reliability, you would either have to pick one of the two types of reliability, or you would have to use an aggregate measure of its reliability. Let us consider these possibilities in turn.

Taking just one of the two measures by itself—for example, the reliability of positive results—and regarding it as indicative of the epistemic quality of a given method of inquiry is highly problematic, and not only because it seems arbitrary as long as you do not yet know what the result is going to be. A type of inquiry can only be relevant with regard to the question whether *p* if it gives a signal in support of *p* under certain circumstances and does *not* do so under certain other circumstances. Taken by itself, a high reliability of positive results is easy, almost trivial to come by—for example, by designing a method that almost never produces positive results (only in obvious cases). For instance, the method of only identifying a tree as an oak tree when it is actually now carrying acorns has a high reliability of positive results, but it is not an overall commendable epistemic strategy in most situations. The same holds for the reliability of negative results if taken by itself. The method of always and only identifying a given tree as an oak tree if it is not carrying fir cones yields highly reliable negative results, but it is obviously epistemically deficient.

One might be tempted to conclude that what we obviously want is some kind of combined reliability. Let the epistemic quality be measured by the reliability of all results, no matter whether positive or negative. I will call this measure, i.e. the proportion of truth among the positive and negative results taken together, a method's "overall reliability." However, overall reliability in this sense can be very misleading about the epistemic quality of a method. In many contexts, it is the sensitivity to a particular kind of case that is relevant to the epistemic quality of a method of inquiry, and not its reliability across all kinds of situations. This point is best demonstrated by considering a concrete example.

Consider, again, the example of the rapid HIV test, this time performed on the general population of Germany. The numbers I will quote in what follows are simplified but realistic estimates of the likely results of such an application,[9] based on the known sensitivity and specificity of such tests and current estimates of the prevalence of HIV in the German population, according to which an estimated 40,000 have the virus. Of these, virtually all would be expected to test positive due to the very high sensitivity of the test. However, due to the very large number of healthy people tested (and due to the fact that the test's specificity is almost, but not quite as good as its sensitivity), there would probably be around 239,880 cases of false positive results, although an overwhelming number of 79,720,120 would correctly receive a negative result. With 239,880 false results out of 80 million, the overall reliability of the test calculates as 0.9970.

Now compare this to the following fictitious placebo test, which simply always returns a negative test result, come what may. Applied to the German population, this would result in 40,000 false negative results (viz., all the results of the people who are actually carrying the virus), with the remaining 79,960,000 correctly receiving a negative result. With only 40,000 false results, this "test" would obviously have a higher overall reliability, namely 0.9995, than the actual rapid HIV test. Nonetheless, it is intuitively clear that the placebo "test" is epistemically completely useless, whereas the rapid HIV test is not.

We are therefore confronted with an example in which a method with lower overall reliability is epistemically preferable to an alternative with higher overall reliability. The insight can be generalized by identifying kinds of reasons that can under certain circumstances justify a preference for a method with lower overall reliability. Such a reason exists, for example, when you aim to avoid both types of mistakes (false negatives and false positives), but one type is particularly important to avoid. (Identifying individual cases of an infectious disease is a case in point, if the aim of limiting the spread of the disease is of particular importance.)

Another example of such a reason is given when you are searching for a phenomenon of very low prevalence. Think of a researcher collecting specimens in the field. If the target subspecies she is investigating is very rare, she has to take particular care in avoiding false

[9] Here, as in the earlier passage about the rapid HIV test, I am calculating for a hypothetical test with a sensitivity of 100 % and a specificity of 99.7 %. A recent WHO report gives these numbers as results of a laboratory assessment of one commercially available rapid diagnostic test (WHO 2015, 23). Other such tests were assessed with very similar results (with slightly worse figures for specificity in some cases, slightly better ones in others) (ibid., 23-25). My hypothetical test by far exceeds the minimum requirements for HIV tests set by the WHO/UNAIDS, which are 99 % sensitivity and 98 % specificity.

negatives when she picks up each individual and decides whether or not it belongs to the subspecies in question and should therefore be collected. (On closer reflection, this second type of reason is thus a special case of the former.) The "method" of just never putting anything into the jar might very well have higher overall reliability than her careful scrutiny, but it is also obviously unsuited to get any investigation started. The example helps to point out that there may be a variety of reasons why you want to avoid one particular type of mistake. In particular, the possible extra-scientific consequences are not the only potential source of such a concern. Even if you are abstracting from such consequences, as the specimen-collecting researcher might very well be, differences in seriousness between false positive and false negative mistakes can be relevant and characteristic of the kind of inquiry that you are engaged with.

A preliminary conclusion from our reflections on reliability so far can be formulated as follows: A meaningful assessment of the epistemic quality of a given method must at the very least account for *both* the reliability of positive results *and* the reliability of negative results (at the same time, but separately). Note that these two types of reliability are in systematic tension: You can typically increase one at the cost of the other without increasing the data input or otherwise investing more effort. Therefore, every methodological decision involves a trade-off between them.

Can the two dimensions we have identified so far (the reliability of positive results and the reliability of negative results) *in combination* be considered sufficient for characterizing the epistemic quality of a method or type of inquiry? To see that they cannot, it is now important to move away from examples of standardized testing procedures.

Let us now consider types of inquiry for which one can draw a line between cases where the method produces a *negative result* and cases where there is *no result at all*. Many standardized procedures always have a result that is regarded as either positive or negative. But this is not so for many cases of scientific inquiry that do not just consist of applying a standard procedure. They often end in an inconclusive result, or lead to the resolve that more work is needed before a result of the inquiry can be declared. In practice, the line between negative results and no-result-cases may be fuzzy in some cases, but for many actual types of scientific inquiry, established and meaningful uses of this distinction exist. Methods or types of inquiry for which this holds differ not only with regard to how reliable their positive results are and how reliable their negative results are, but also with regard to how likely they are to produce any result at all, given a certain amount of effort expended.

This enables us to identify an important additional dimension that is relevant to how good a method is at getting at the truth, which I shall call the investigation's power. Power in this sense is the rate at which a method or type of inquiry generates definitive results, given a certain amount of effort and resources.[10]

[10] "Power" also is the name that Alvin Goldman (1992: 195) uses to identify a similar characteristic of a method, namely the rate at which it produces *true* results. Goldman's category of power thus already combines what I propose to call power with reliability.

A little reflection makes clear that even the consideration of both the reliability of positive results and the reliability of negative results cannot result in a significant measure of epistemic quality unless power is taken into account as well. In the absence of constraints on power, high reliability of both types of results usually comes cheap. For example, if we lowered the figure for the highest reasonable significance level that justifies the label "statistically significant result" from .05 to .01, we would have much more reliable results. But we would also have dramatically fewer results. The aim of getting at the truth implies not only that we want reliable results, but also that we want results.

The resources that we are able to dedicate to inquiry are always limited, which is why a method's power crucially matters to our epistemic aims. Similar to the two types of reliability, methodological choices also involve a trade-off between *both* types of reliability, on the one hand, and the power of the investigation, on the other. As an illustration, compare reliance on purely observational data in medicine with exclusive reliance on large randomized controlled trials (RCTs). You may want to ask: Which of the two strategies is more properly geared towards the truth? But this question as such cannot be answered, because both trade off power and reliability in different ways. RCTs (let us grant) are more reliable both in positive and negative results, but much less powerful: Observational data give us more results for the same amount of effort and resources. So the question which one of them is more properly geared towards the truth makes no sense, because both reliability and power are important dimensions of what it means for a method to be geared towards the truth. As an aside, one of the problems with the hierarchy of evidence as it is advocated by the movement for "Evidence Based Medicine" (EBM) seems to be that it gives all the attention and priority to reliability and none at all to power.[11]

What is desirable in a method of inquiry (from an epistemic perspective) can only be captured by considering all three dimensions: the reliability of positive results, the reliability of negative results, and the method's power. For each method, these three magnitudes form a triple that I will call the inquiry's *distribution of inductive risks*, or "DIR," for short.

DIR = ⟨reliability of positive results, reliability of negative results, power⟩.

This label takes up the talk of different types of inductive risk that was first introduced by Carl Hempel (1965: 91-92) and recently revived by Heather Douglas (2000, 2009). It extends the concept of inductive risk to include not only the risk of false negative mistakes and the risk of false positive mistakes but also the risk of ending up without any result at all.

If DIR is needed to provide a description of those features of a method that are relevant to its epistemic quality, important consequences follow. First of all, there is no obvious strategy of "maximizing" DIR or optimizing it in any other straightforward way. DIR is not a scalar quantity but a three-dimensional vector. What is more, the three dimensions of the vector are antagonistic to each other in the sense that each of them alone can easily be increased at the cost of one or both of the others, so that any methodological choice involves a trade-off between the three dimensions. Consequently, there can be no linear ranking of methods according to their purely

[11] For critical methodological views on EBM, see Worrall 2002, Borgerson 2009, Goldenberg 2006.

epistemic quality. Since all three dimensions of DIR are geared towards the aim of getting at the truth, methodological choices therefore turn out to be underdetermined by truth as the aim of inquiry.

Another important consequence can be recognized when we turn to the question how the appropriate DIR for a particular case of inquiry can or should be determined. Is there a non-arbitrary way in which we may speak of the "correct" DIR? Early statisticians like Abraham Wald or C. West Churchman faced a similar question when they discussed the issue of how strong the statistical evidence must be in order to warrant the acceptance (or, respectively, the rejection) of a hypothesis. They were in agreement that the science of statistics cannot answer the question. Rather, the correct answer depends on what is at stake in the particular case (Wald 1942: 40-1; Churchman 1948: ch. 15). This insight about statistics can be generalized: A non-arbitrary sense of a correct DIR can only be determined as a function of certain value judgments.[12] The value judgments in question are judgments about *how* valuable, in the specific circumstances of each particular investigation, a correct positive result would be as compared to the state of continued ignorance, and similarly, how valuable a correct negative result would be, how bad a false negative result would be and how bad a false positive result would be. Without answers to these questions, meaningful comparisons between different distributions of inductive risks are underdetermined.[13]

5. Methodological Standards and the Distribution of Inductive Risks

While the only non-arbitrary sense of a correct DIR makes it dependent on value judgments, making the appropriate trade-offs and determining DIR for each and every investigation is in scientific practice not left to individual researchers and *their* value judgments. Instead, DIR is heavily constrained by the respective research community's methodological standards. As an illustration, consider again the methodological rule that you need a causal hypothesis in biosafety studies on GMO. By adopting this rule, you can lower the risk of getting false positive results (which in this case would mean affirming a positive correlation between the presence of the GMO and some undesirable event in the environment where in reality there is no causal connection between the two). For if your search for correlations is always guided by a causal hypothesis, you are less likely to be misled into positing some effect—concerning, say, the survival rates of one of the insect species in the environment—that later turns out to be an artefact

[12] I explore the exact nature of this dependence in Wilholt 2013.

[13] This line of reasoning for the value-dependence of DIR is essentially the same as the one that Churchman's student Richard Rudner already used to argue that "The Scientist *qua* Scientist Makes Value Judgments," in his essay of the same title (1953). Owing in great part to the work of Heather Douglas (2000, 2009), who has rediscovered the argument's importance and explored its consequences and ramifications, it now again plays a central role in debates over science and values—rightly so, I believe (cf. Wilholt 2009, 2013). Note that my version of the argument in the present paper does not start from particular assumptions about the moral responsibilities of scientists. Rather, the core argument is that there can be no linear measure of the quality of a method that derives only from the single aim of finding truth and avoiding falsehood. Methodological choices are therefore underdetermined by this aim. In *consequence*, every non-arbitrary methodological choice involves value judgments, and thus the (moral, social, political) responsibility of scientists is the only thing that can fill the gap which a "purely epistemic" perspective on methodology must necessarily leave.

of the study design or even a random pattern in your dataset. At the same time, the rule increases the risk of false negative results (i.e. the risk of missing an effect that really is there). This is so because studies that are *not* restricted by the proposed rule and that simply look for all sorts of changes that are statistically correlated with the presence of GMOs will arguably at least occasionally hit upon an environmental effect that is real but unexpected in the sense that we had no causal conjecture about it beforehand. (Both our causal understanding of biological mechanisms and our creativity in thinking about them are imperfect. Therefore, if it *seems* to us that our present understanding of the mechanisms provides no clue to a connection, it does not mean that there couldn't be one.) The proposed methodological rule would therefore go along with shifting some of the inductive risk from the false positives to the false negatives. To return to another one of our earlier examples, the selective publication of results of clinical trials has been suspected of leading to a situation in which the body of published research literature "overestimate[s] the benefits of an intervention" (Chan et al. 2004: 2457). Negative results regarding the efficacy of novel interventions (as compared to standard therapy) are particularly likely to fall under the carpet as they are often not published and sometimes even cloaked in secrecy. The introduction of clinical trial repositories as a standard of the biomedical community makes it less likely that negative results fall into oblivion. Due to obligatory registration in publicly accessible repositories, negative outcomes too can become known to all researchers who take an interest in the respective studies, and this knowledge can then spread through the community. The rule thus limits the risk of false negative results, if under a "result" we choose in this case to understand the opinion that the research community will ultimately arrive at on the basis of the available information on concluded trials.

In face of the fact that methodological standards impose constraints on DIR, one could still attempt to uphold the view that the role of research communities in providing and maintaining such standards is merely accidental—auxiliary, as it were. In order to do so, one would have to argue that the standards save individual researchers the trouble of making the appropriate value judgments and finding the methodological choices that affect DIR accordingly by themselves. The obvious shortcoming of this line of reasoning is that it presupposes it to be somehow determined and uncontroversial what the appropriate value judgments are. While members of research communities tend to share *some* interests, it is unlikely that they would generally arrive at the same value judgments about the benefits of true results and the seriousness of mistakes. Methodological standards thus not only provide a service that makes the job of balancing out inductive risks easier. They also *harmonize* DIR within research communities. The value judgments implicit in the constraints on DIR that the standards set are in a sense binding for the community's members—not in the sense that the researchers are individually committed to endorsing the value judgments, but in the sense that they have to act as if they endorsed them when they perform certain research-related actions.

Should this imposition of a research community on its members perhaps be criticized or even opposed? On the contrary, I think it is hardly imaginable how a collective cognitive enterprise could work without it. In order to enable and maintain epistemic trust *within* a research

collective, it is necessary to *co-ordinate* the DIR that is deemed acceptable in its research activities. At first sight it might seem a possible arrangement to let every scientist freely choose his or her own DIR in accordance with his or her own value judgments about how serious each particular kind of mistake would be. But such individual decisions would be extremely cumbersome to track and take account of by peers. Not only can value judgments vary considerably from individual to individual, it is also usually difficult to guess another person's value judgments on a given subject matter. But as we have seen, DIR is the decisive characteristic for assessing the trustworthiness of a given piece of scientific information for a given purpose. Methodological standards that put constraints on DIR thus make such an assessment of another researcher's results possible without having to take a guess at his or her value judgments.

That said, there is perhaps an additional reason why it is not at all objectionable if the value judgments that determine the limits on acceptable inductive risks are made by a collective rather than by each individual. Collectives or communities might simply be better placed to make such judgments, especially if they incorporate a diversity of perspectives and value outlooks. This may make them less prone to oversights and biases than an individual. While I agree that collective deliberation on value judgments can have these positive effects, my argument in this paper does not rely on them. My main argument why constraints on DIR and the required implicit value judgment underlying them need to be determined on the community level is that harmonization is required in order to facilitate and maintain epistemic trust within research communities.

In conducting her scientific work as a member of a particular research community, a researcher implicitly commits herself to respecting its methodological standards and thereby to working within the limitations on acceptable distributions of inductive risks that these imply. In so far as this serves to harmonize DIR within the community, the precise constraints imposed by methodological standards turn out to be at least in part conventional in the following sense. The purpose of enabling trust within a community by harmonizing accepted distributions of inductive risks could presumably be served by a whole range of diverging determinations of what the ideal DIR is. With regard to the aim of facilitating reliable assessments of the trustworthiness of other researchers' results, it is crucial that everyone within the community sticks to *the same* standards and thus the same limitations on DIR, but not *which* particular DIR it is that is set as an ideal. The standards provide a solution to a problem of coordination and are thus conventions. They are only *partly* conventional because they do not *only* serve the purpose of harmonizing DIR within the community. They also represent the research community's collective attempt to find the *right* balance between power and the two types of reliability. In that sense, they also represent an implicit consensus (or at least an implicit compromise position) of the community with regard to the question of how valuable the benefits of correct results and how grave the negative consequences of mistakes typically are for the kinds of research procedures that are subject to the standards at issue.

6. Conclusions

Research communities are bound together by methodological standards that shape DIR in their respective area of research. These standards serve as a solution to a problem of coordination and are thus irreducibly social—that is to say, communities play an *essential* role in shaping DIR. Without this role, the kind of collective epistemic enterprise that we call science, with its high degree of reliance on each others' results and its fast spread of novel information through the community, could arguably not get off the ground.

At the same time, DIR is decisive for assessing the trustworthiness of a piece of scientific information. It follows that every reasonable assessment of trustworthiness with regard to scientific results must attend to research communities. In placing our trust in a scientific result, we in part place our trust in the respective research community and its methodological standards. We trust that this community has set the limitations on the DIR in a suitable way, considering the risks and benefits of false and correct results appropriately. Ideally, someone investing trust in a scientific result would, of course, like to see the research communities weigh the risks and benefits in a way that is at least roughly in line with her own value judgments on the matter. (But note that a different, more circumspect kind of reliance in scientific results remains possible as long as the community determines the DIR in a *predictable* way, even if this diverges from how we ourselves as "users" of the information would have done it.) A part of our trust is, of course, also placed in the individual researcher (or collaborative group) generating the respective piece of scientific information.[14] A research project cannot be executed like an algorithm after all—for all the methodological conventions, skill and competence are of no less crucial importance in scientific research. But in the kind of collective enterprise that we call science, every individual researcher or collaborative group is also part of a research community whose standards impose decisive constraints on methodological choices and which therefore always has a share in the production of the results.

Does the social diffusion of trustworthiness to the level of research communities bear any wider significance for the social epistemology of scientific research? For a start, it means that philosophers of science should take an interest in conventional methodological standards and the various ways they come about. Some standards take the form of tacit rules that proliferate certain entrenched practices and are habitually passed on in professional training. Some originate as explicit recommendations of task forces or peer review panels initiated by professional associations or government agencies. Some are based on unilateral decisions by editors of influential journals. Some may even be inscribed in widely used research tools, such as statistics software packages. In investigating the standards and the ways they arise, philosophers should keep in mind that they inevitably contain implicit value judgments on the benefits of correct results and the disadvantages of mistakes.

[14] I do therefore not wish to dispute that trust in individual colleagues still plays an important role in present day science, as recent writers have emphasized (Shapin 2008, Frost-Arnold 2013). But the potential of trust directed at individuals alone to explain and serve as a basis of the division of cognitive labor in the sciences is very limited in light of the phenomena I have discussed. It has to be supplemented by trust directed at collective bodies.

The latter point leads to a more far-reaching conclusion. When we want to assess the relative merits of the collective methodological decisions of a research community, the aforementioned implicit value-judgments will have to come under scrutiny, too. How well do they reflect the epistemic role that the research community fulfils within wider society? This question cannot be answered with respect to just the single aim of getting at the truth. We have seen that it is a deeply problematic take on epistemic quality management to consider the reliability of a method to provide a simple, one-dimensional measure of how good a method is at getting at the truth. Power and both types of reliability are all aspects of aiming at the truth, but aspects that need to be traded off against one another. The question what the right balance between them is involves types of valuations that cannot themselves be derived from the aim of truth alone. Call an evaluative concept "purely epistemic" if it involves only evaluations with regard to the aim of acquiring true beliefs and avoiding false ones. Then, what we have observed means that there can be no purely epistemic linear measure of the quality of a method. A purely epistemic perspective on methodology is always incomplete.

Acknowledgments

This paper was first presented at the March 2011 conference on *Collective Epistemology*, organized by Miranda Fricker and Michael Brady, from which this volume originates. Different parts of it were also presented at the *SPSP* conference in Exeter in June 2011, at the *EPSA* in Athens in October 2011, at a conference on *The Social Relevance of Philosophy of Science* at the Center for Interdisciplinary Research in Bielefeld in June 2012 and at the *GAP.8* in Konstanz in September 2012. I am grateful to the organizers and participants of all these events, and particularly to my co-panelists at Exeter and Athens: Justin Biddle, Bryce Huebner, Rebecca Kukla and Eric Winsberg.

References

The ATLAS Collaboration, G. Aad, E. Abat et al. (2008): "The ATLAS Experiment at the CERN Large Hadron Collider", *Journal of Instrumentation* 3, S08003.

Borgerson, Kirstin (2009): "Valuing Evidence: Bias and the Evidence Hierarchy of Evidence-Based Medicine", *Perspectives in Biology and Medicine* 52 (2), 218-233.

Broer, Inge (2012): "Divergierende naturwissenschaftliche Bewertung der Grünen Gentechnik: Grundlagen biologischer Risikoanalyse", in: *Grüne Gentechnik: Zwischen Forschungsfreiheit und Anwendungsrisiko*, ed. by Herwig Grimm & Stephan Schleissing, Baden-Baden: Nomos, 81-92.

Chan, An-Wen, Asbjørn Hróbjartsson, Mette T. Haahr, Peter C. Gøtzsche and Douglas G. Altman (2004): "Empirical Evidence for Selective Reporting of Outcomes in Randomized Trials: Comparison of Protocols to Published Articles", *Journal of the American Medical Association* 291, 2457-2465.

Churchman, C. West (1948): *Theory of Experimental Inference*, New York: Macmillan.

Douglas, Heather (2000): "Inductive Risk and Values in Science", *Philosophy of Science* 67, 559-579.

Douglas, Heather (2009): *Science, Policy, and the Value-Free Ideal*, Pittsburgh: University of Pittsburgh Press.

Frost-Arnold, Karen (2013): "Moral Trust and Scientific Collaboration", *Studies in History and Philosophy of Science* 44 (3), 301-310.

Goldberg, Sanford C. (2006): "The Social Diffusion of Warrant and Rationality", *The Southern Journal of Philosophy* 44, 118-138.

Goldenberg, Maya J. (2006): "On Evidence and Evidence-Based Medicine: Lessons from the Philosophy of Science", *Social Science and Medicine* 62 (11), 2621-2632.

Goldman, Alvin I. (1992): *Liaisons: Philosophy Meets the Cognitive and Social Sciences*, Cambridge, MA: MIT Press.

The GUSTO Investigators (1993): "An International Randomized Trial Comparing Four Thrombolytic Strategies for Acute Myocardial Infarction", *New England Journal of Medicine* 329, 673-682.

Hardwig, John (1985): "Epistemic Dependence", *The Journal of Philosophy* 82 (7), 335-349.

Hempel, Carl G. (1965): *Aspects of Scientific Explanation*, New York: Free Press.

Kukla, Rebecca (2012): "'Author TBD': Radical Collaboration in Contemporary Biomedical Research", *Philosophy of Science* 79 (5), 845-858.

Longino, Helen (1990): *Science as Social Knowledge. Values and Objectivity in Scientific Inquiry*, Princeton: Princeton University Press.

Longino, Helen (2002): *The Fate of Knowledge*, Princeton: Princeton University Press.

National Science Board (2012): *Science and Engineering Indicators 2012*, Arlington: National Science Foundation, http://www.nsf.gov/statistics/seind12/pdf/seind12.pdf. (Accessed 25 March 2015).

NTP (2001): *National Toxicology Program's report of the endocrine disruptors low-dose peer review*, National Toxicology Program, U.S. Department of Health and Human Services, http://ntp.niehs.nih.gov/ntp/htdocs/liason/LowDosePeerFinalRpt.pdf. (Accessed 25 March 2015).

Rudner, Richard (1953): "The Scientist *qua* Scientist Makes Value Judgments", *Philosophy of Science* 20 (1), 1-6.

Shapin, Steven (2008): *The Scientific Life: A Moral History of a Late Modern Vocation*, Chicago: University of Chicago Press.

Surowiecki, James (2005): *The Wisdom of Crowds: Why the Many Are Smarter Than the Few*, London: Abacus.

Thagard, Paul (1997): "Collaborative Knowledge", *Noûs* 31 (2), 242-261.

Tollefsen, Deborah (2007): "Group Testimony", *Social Epistemology* 21 (3), 299-311.

Wald, Abraham (1942): *On the Principles of Statistical Inference* (= *Notre Dame Mathematical Lectures* 1). Notre Dame, IN: University of Notre Dame.

WHO (2015): *HIV Assays: Laboratory Performance and Other Operational Characteristics: Rapid Diagnostic Tests (Combined Detection of HIV-1/2 Antibodies and Discriminatory Detection of HIV-1 and HIV-2 Antibodies), Report 18*, Geneva: World Health Organization, http://www.who.int/diagnostics_laboratory/publications/evaluations/en/. (Accessed 25 March 2015).

Wilholt, Torsten (2009): "Bias and Values in Scientific Research", *Studies in History and Philosophy of Science* 40, pp. 92-101.

Wilholt, Torsten (2013): "Epistemic Trust in Science", *The British Journal for the Philosophy of Science*, 64 (2), 233-253.

Winsberg, Eric, Bryce Huebner and Rebecca Kukla: "Accountability and Values in Radically Collaborative Research", *Studies in History and Philosophy of Science* 46, 16-23.

Worrall, John (2002): "What Evidence in Evidence-Based Medicine?", *Philosophy of Science* 69 (Proceedings, S3), S316-S330.

Wray, K. Brad (2002): "The Epistemic Significance of Collaborative Research", *Philosophy of Science* 69 (1), 150-168.

Wray, K. Brad (2006): "Scientific Authorship in the Age of Collaborative Research", *Studies in History and Philosophy of Science* 37 (3), 505-514.