Bias and values in scientific research

Torsten Wilholt

Department of Philosophy
Bielefeld University
P.O. Box 100131
D-33501 Bielefeld, Germany

twilholt@uni-bielefeld.de

Manuscript version. The original article is forthcoming in *Studies in History and Philosophy of Science* and will be available at http://www.sciencedirect.com/science/journal/00393681.

Abstract

When interests and preferences of researchers or their sponsors cause bias in experimental design, data interpretation or dissemination of research results, we normally think of it as an epistemic shortcoming. But as a result of the debate on science and values, the idea that all "extra-scientific" influences on research could be singled out and separated from pure science is now widely believed to be an illusion. I argue that nonetheless, there are cases in which research is rightfully regarded as epistemologically deficient due to the influence of preferences on its outcomes. I present examples from biomedical research and offer an analysis in terms of social epistemology.

Keywords: science and values, bias, social epistemology

Bias is becoming increasingly recognized as a serious problem in many areas of scientific research. Of particular concern are cases in which research results seem directly to reflect the preferences and interests of certain actors involved in the research process. Troubling examples of this have been identified, especially in privately funded research and in policy-related areas.

Intuitively (and traditionally) it seems clear that the suggested kind of bias constitutes outright epistemic failure. But philosophers of science have begun to realize that the ideal of pure and value-free science is at best just that—an ideal—and that all scientific practice involves all kinds of value-judgments. While some

philosophers have sought to distinguish acceptable from unacceptable influences of values on science, efforts to draw this distinction in a principled way have proven immensely difficult (see sec. 5). So why should not some values that inform scientific research be, e.g., shareholder values?

My primary aim in this paper is to describe and define the suggested kind of bias in a way that allows us to characterize it as an epistemic shortcoming of the research in question. I will end up arguing that one need not deny the inevitable value-ladenness of science in order to mark certain cases of bias as being scientifically unacceptable.

Note that my aim is not to analyze the *concept* of bias. There are many widely differing uses of "bias" both within science and within philosophy—enough to suggest that the word is polysemic (cf. Gluud 2006, Goldman 1999, sec. 8.3, Resnik 2000). I am interested in a certain *phenomenon*, which I will introduce with the help of examples in the following section and try to characterize provisionally.

1. Preference bias

In the context of science and values, a phenomenon that I will call *preference bias* is of particular interest. It occurs when a research result unduly reflects the researchers' preference for it over other possible results. (Note that this is a special kind of bias; the term 'bias' is also often applied to cases of systematic error that need have nothing to do with investigators' preferences for one result or another. A classic example is the kind of bias in clinical trials introduced by incomplete randomization, which tends to re-confirm, if anything, the researchers' preconceived *beliefs* rather than their preferences. Cf. Gluud 2006, pp. 494-495) One important caveat is that preference bias should be distinguished from outright falsification or fabrication of results. Preference bias works in a more subtle way, by increasing the likelihood of the preferred outcome rather than by bluntly fabricating it.

Before I turn to the task of giving a more precise and satisfactory characterization of preference bias, I would like to present some examples of the phenomenon that have recently raised concern in the biomedical literature. They illustrate the variety of mechanisms by which the researchers' preferences can come to exert a problematic kind of influence on the research result. Particular cases of preference bias are almost always controversial. For the concerns of this paper, it is inessential whether the controversies over either of the following examples can be considered resolved. What interests us here is the *charge* that preference bias has compromised the research in question, and the philosophical

problem of how best to characterize the kind of shortcoming that is implied by such a charge.

Bisphenol A is used as a monomer in polycarbonate plastic and has been related to cancer and other health problems. Its toxicity is associated with its similarity to human estrogen. A controversial issue is the health risk of exposure to low doses. Biomedical scientists Frederick vom Saal and Claude Hughes noted that 90% of government-funded experiments on low-dose exposure to bisphenol A reported significant effects, while not a single industry-funded experimental study did so (vom Saal & Hughes 2005). What is more, they found that some industry studies used a strain of rat (the CD-SD strain) that is particularly insensitive to any estrogen. Two industry studies initially included positive control groups that were exposed to the well-characterized estrogenic drug DES. The positive and negative controls failed to exhibit significant differences; this could have alerted the investigators to the unsuitability of their strain of experimental animal. Instead, in both cases the researchers chose to ignore this outcome and not to mention the positive control in their publications. Subsequent industry-funded studies simply omitted a positive control (vom Saal & Hughes 2005, pp. 928-929, vom Saal & Welshons 2005, p. 52). If these characterizations are adequate, the respective industry-funded studies on bisphenol A may be said to suffer from *biased experimental design*. The design of the studies made it unlikely to detect any effects.

Biased experimental design may also afflict some randomized drug trials. At least, that is one plausible explanation of the frequently observed phenomenon that results of trials are significantly more favorable towards experimental interventions when they are funded by for-profit organizations (see e.g. Kjaergard & Als-Nielsen 2002; cf. Bekelman et al. 2003 and Lexchin et al. 2003 for systematic reviews). One factor contributing to this effect may be due to the choice of control intervention. Helle Johansen and Peter Gøtzsche (1999) have criticized industry-funded studies of the antifungal agent fluconazole for unfairly comparing this intravenously administered drug with a control intervention of nystatin that was orally administered and thus relatively poorly absorbed. Benjamin Djulbegovic and colleagues (2000) investigated 136 published randomized trials on patients with multiple myeloma and found that most industry sponsored studies (which were on the whole much more likely to end up favoring the experimental treatment) compared the experimental intervention to a placebo or to no therapy, while most publicly sponsored studies by far used an active (standard) control therapy as comparator. The funding effect may thus be in part caused by a particular kind of biased experimental design, *viz*. use of a substandard comparison.

Preference bias can also be found *after* experiments, trials or studies have been performed in accordance with a certain design. Another toxic substance used for polymerization is vinyl chloride. Its toxicity is in many respects well established, including its carcinogenic agency with regard to cancer of the liver. However, its link with other kinds of cancer became subject to protracted dispute, despite accumulating evidence for its linkage with, among others, cancer of the brain (cf. Sass et al. 2005, Markowitz & Rosner 2002, ch. 7). In 1988, a review of epidemiological data from several European and North American studies on men occupationally exposed to vinyl chloride reported an increased occurrence of brain cancer, the standardized mortality ratio (SMR, i.e. the ratio of observed deaths to expected deaths times 100) being 148. In his interpretation of this result, Richard Doll, the author of the review, chose one small contributing study and subtracted the four cases of brain cancer death reported by it from the total in his survey, on the grounds that this study was itself the origin of the hypothesis that vinyl chloride might cause brain cancer and therefore this hypothesis should be tested by the remaining data, and also because it was "not a cohort study" (Doll 1988, p. 70). Note that no similar operation is performed during the interpretation of the many other results of the survey. After this operation, there is still an excess of brain cancer (SMR = 131), but the total numbers are now too small to be statistically significant. Doll concluded: "No positive evidence of a hazard of […] any type of cancer other than angiosarcoma of the liver has been found […]". Three years later, another study found excess deaths from cancer of the brain and the central nervous system "confirmed," reporting a SMR of 180 (Wong et al. 1990). But soon after, two of its authors recanted and re-interpreted their findings as a possible product of "diagnostic sensitivity bias," due to "more complete reporting and/or diagnoses of brain tumors in employees of large corporations than in the general population" (Wong and Whorton 1993). A decade later, a third study again reported excess brain cancer deaths among exposed workers (SMR 142, and SMR 177 among subjects with the longest work history), but when they turned to an interpretation of their findings regarding mortality from brain cancer, the authors commented that "its relation with exposure to vinyl chloride remains unclear" (Mundt et al. 2000, p. 774). Jennifer Sass and colleagues (2005) have concluded that the evidence for the linkage of vinyl chloride with brain cancer has been consistently downplayed. They also point to the fact that all three studies were commissioned by the chemical industry. There is evidence that the remarkable recantation of Wong and Whorton was prompted by members of the Vinyl Chloride Panel of the Chemical Manufacturer's Association (Markowitz & Rosner 2002, pp. 229-230). If this characterization of the story is accurate, then the example provides ample illustration of how *biased interpretation of outcomes* is possible:

Outcomes can be declared irrelevant, they can be attributed to a speculative alternative cause, or they can be declared insufficient for the validation of a decisive conclusion.

One additional kind of bias that needs to be addressed is *biased communication and dissemination of results*. Here, the best known phenomenon is publication bias, due to the fact that only a portion of all research results are published. Experiments or trials with significant outcomes are much more likely to be published than such with null or weak results, with the result that the overall picture in the scientific literature (as captured, for example, in literature reviews and meta-analyses) is in some cases significantly distorted (see Song et al. 2000 for review and comprehensive analysis). Editorial practices and journal referee decisions are thought to be partially responsible for publication bias (*ibid.*, pp. 28-30). But also the reluctance of investigators or their sponsors to see unwelcome results published is known to contribute to the effect, in which case we are confronted with clear cases of preference bias (*ibid.*, pp. 30-32). There is evidence to suggest that in biomedical research, subtle and not so subtle mechanisms frequently prevent or delay dissemination of results when they run counter to the business interests of sponsors (Blumenthal et al. 1997). Recently, evidence has solidified that selective publication does not only affect complete studies, but that also within studies (which typically result in a large set of outcomes), outcomes are often reported selectively, depending on the nature of each respective result. Cautiously, the authors of the respective investigation note that as a result, the published literature may "overestimate the benefits of an intervention" (Chan et al. 2004, p. 2457).

These cases give only a small sample of the many ways in which judgements and decisions of scientists can be affected, and they can only hint at the larger consequences for science (for a wealth of cases, see Krimsky 2003, Resnik 2007). Nevertheless, the examples given suffice to illustrate how preferences of investigators (and indirectly also preferences of research sponsors) can exert their influence on research outcomes at several major steps of the research process. In the following section, I will consider a preliminary and intuitive way of analyzing such influences as biases. Although the analysis will under closer scrutiny turn out to be problematic, it is helpful in identifying the common traits of the cases at issue and in approaching the underlying epistemological problems.

2. Preference bias and inductive risk

In every empirical investigation that is designed to test some hypothesis *H*, two kinds of risk can be identified: the risk that the investigation may lead to the

acceptance of *H* while *H* is in fact false, and, conversely, the risk of rejecting *H* when *H* is in fact true. Carl Hempel (1965, pp. 91-92) has coined the term "inductive risk" to cover these two types of risk. It was recognized early on in the development of statistics that in contexts relevant for practical applications, the consequences of the two types of error are typically borne by different parties. Inspired by the terminology of quality control, one of the most important early applications of statistics, it became common to label one type of risk (originally, the risk of falsely rejecting a good product) "producer's risk" and the other "consumer's risk" (see e.g. Pearson 1933).

While a good experiment, study or trial is always designed to lower the total risk of leading to a false conclusion, this risk cannot be minimized beyond a certain limit without increasing the empirical input. However, the crucial insight in our context is that one *type* of inductive risk can usually be *traded off* against the other type without changing the size of the empirical basis, by means of alterations regarding experimental design, data analysis or even dissemination of results. E.g., by their choice of a particular strain of rat, researchers investigating the effects of low-dose exposure to bisphenol A reduced the risk of falsely postulating an effect where there is in fact none, while by the same move they increased the risk of negating effects that really exist.

In all our examples for preference bias, it may plausibly be assumed that the involved actors' preferences have induced them to trade off a decreased producer's risk against an increased consumer's risk. (For cases of publication bias, this presupposes that the relevant sense of acceptance of the hypothesis is understood as acceptance within the community that absorbs the published research literature.) Motivated by this common feature, let us consider the possibility to simply conceive of preference bias as the result of tampering with the balance of inductive risks. According to this idea, preference bias would be regarded as a researcher's failure to be impartial between the two kinds of risk, and allowing her different attitudes with regard to the desirability of a positive or negative result to influence the set-up of the test or even the whole research project in such a way that one type of inductive risk is decreased at the expense of the other's amplification. This analysis could cover manipulations intentionally introduced by the investigator in order to increase the chance of arriving at the desired results, as well as subconscious influences of her preferences on her methodological choices.

3. Inductive risk and the evaluation of outcomes

This tentative analysis of preference bias in terms of inductive risk seems to face a serious problem, though. The analysis under consideration starts from the implicit

assumption that there is a certain correct or impartial balance between the two kinds of inductive risk that exists independent of the researcher's preferences, and that preference bias consists in the deviation from that balance. However, it has long been argued that in cases where the aim is to accept or reject a hypothesis on the basis of evidence, there is no non-arbitrary and convincing way to strike a correct balance of inductive risks independent of certain value-judgments. The point was first argued explicitly by statisticians such as Abraham Wald (1942, pp. 40-41) and C. West Churchman (1948, ch. 15), but came to prominence through a brief paper by Richard Rudner (1953). It remains one of the most forceful arguments for the inevitability of value-judgments within scientific research.

The argument is quite simple and straightforward. No empirical hypothesis is ever completely verified by the data. In order to accept a hypothesis, a scientist must decide which level of inductive confirmation she considers sufficient for acceptance. With regard to the analysis and interpretation of the data, in classical statistics this choice is most obviously present in the selection of a level of significance, but as we have seen, all kinds of decisions at all levels of the design and execution of an empirical investigation have an influence on how strong the evidence will have to be in order to actually lead to the hypothesis' acceptance. The appropriateness of these decisions depends on the minimum level of probability that the investigator chooses to be sufficient for the acceptance of the hypothesis, and there is nothing in the science of statistics or in the logic of inductive reasoning to determine this choice. Instead, Churchman and Rudner argue, the researcher's evaluation of the possible outcomes of the investigation will have to determine the choice. In particular, the evaluation of the two kinds of possible mistakes should make a difference to the minimum degree of confirmation at which the hypothesis becomes acceptable. "How sure we need to be before we accept a hypothesis will depend on how serious a mistake would be." (Rudner 1953, p. 2) Therefore, "the value of any test procedure depends upon a certain function of both the chance of error and the loss." (Churchman 1948, p. 256) That is to say, how good or bad a given empirical investigation suits its purpose is *always* relative to value-judgments regarding the potential loss that each kind of mistake would entail. Such judgments are therefore part and parcel of the researchers' task of designing and selecting test procedures for hypotheses. (Of course, everything that this argument says about accepting hypotheses applies equally, *mutatis mutandis*, to rejecting them.)

One immediate rejoinder to this argument was formulated by Richard Jeffrey (1956). He argued that the scientist's task, properly understood, was not to reject or accept hypotheses, but rather to establish their degree of confirmation in light of

the available empirical evidence. It is only in decisions about whether or not to *act* on the basis of a given hypothesis that value-judgments about the possible consequences of errors come into play. When such decisions are made, the probabilities of the relevant hypotheses' correctness as established by the scientists should by all means be taken into account, so Jeffrey's rejoinder goes on, but this kind of decision-making will typically fall outside of the scientists' own area of activity.

One rather obvious shortcoming of Jeffrey's response is its mismatch with reality, in particular with regard to the claim that scientists do not, or at any rate should not accept or reject hypotheses but only assign probabilities. Another, more rigorous point was pointed out by Heather Douglas (2000), whose work was vital for bringing the argument of Rudner and others back into the current debate about science and values. While Jeffrey's rejoinder assumes that the need to evaluate the consequences would only arise at the end of the research process and only if the scientist tried to sum up her results in form of an accepted or rejected hypothesis, scientists in fact "also take inductive risks in stages of science before acceptance or rejection of theories, thus considering risks never brought to the light of public decision-making." (Douglas 2000, p. 563) Douglas has shown that there are inductive risks involved in methodological choices, in evidence characterization and in the interpretation of results, and thus in all kinds of decisions that do patently fall within the scientist's own area of work. When finally the hypothesis can be considered in light of the investigation's results, the scientist will already have struck a particular balance between the two types of inductive risk on several occasions within the research process, and she will arguably have had to evaluate the consequences of errors in order to do so.

So where does this all leave us with respect to our characterization of preference bias? It seems that as long as there is no principled way to distinguish the preferences among possible outcomes referred to in our definition of preference bias from the kinds of evaluations of possible outcomes that are always and inevitably present in scientific research according to Douglas, Rudner and others—and I can see no such way—, what we have called "preference bias" will simply be part of the scientific condition. In other words, it is prima facie not possible to distinguish cases of diverging judgments concerning the evaluation of consequences from cases of different preference biases as defined above. As it stands, this conclusion would make a big difference for the criticism of biased research. The controversial aspects of preference bias would seem to arise on the level of value-judgments. The researchers who chose to use a strain of rat that is particularly insensitive to estrogen in order to study the effects of low dose

exposure to bisphenol A did so as a result of their particular evaluation of the consequences of the possible outcomes of their study. Their evaluation presumably was that the possible overregulation that might have resulted from a false positive outcome of their studies would have been a particularly dreadful outcome. It seems to follow that we could only accuse them of having made a mistake or having applied an improper test procedure in the same sense in which we might also be prepared to say that someone makes a mistake or acts improperly by not holding the same values as we do.[1] Bias-talk would thus be revealed as primarily involving the charge of a *moral* shortcoming. This result will surprise everyone who shares my intuition that whether or not the cases described in section 1 display epistemological shortcomings is *not* relative to a set of personal value-judgments. Of course, we all know that intuitions can be deceptive. Should we simply learn to live with the counter-intuitive moral analysis of bias?

4. The ideal of purity

There is one sense in which presumably everyone would agree that the shortcomings of cases of bias are relative to a value; they are at least relative to whichever value is ascribed to replacing ignorance with true belief. A possibility to save the intuition appealed to at the end of the last section might therefore be to describe the treatment of inductive risk as relative to this value and *only* this value. A response to the challenge of Churchman and Rudner along these lines was attempted by Issac Levi in a series of articles (1960, 1961, 1962). Slightly adapting one of his ideas (1962, pp. 55-56), one can develop the following Bayesian consideration. A researcher who is trying to design a test procedure for hypothesis H and who pursues the sole aim of replacing ignorance with true belief should ideally construct the procedure in such a way that it will select among the theoretical options as would a rational agent acting in accordance with the following utility matrix:

	H is in fact true	H is in fact false
accept H	1	0
reject H	0	1
suspend judgment	k	k

[1] Incidentally, note that this point of view comes close to a line of defense that is sometimes used by those accused of bias. For example, in an article written to defend industrial science against the charge of continued bias in the case of vinyl chloride, two industrial researchers make the appeal that "[o]f course, all scientists have biases", invoking an open letter in which fifteen past presidents of the Society of Toxicology endorse that statement (Barrow and Conrad 2006, p. 154).

The fact that the utilities for both correct outcomes (accepting H when H is true and rejecting H when H is false) are equal and positive reflects the assumption that the respective investigation is geared to finding truth, and nothing else. For the same reason, zero utility is assigned to both kinds of error. It is assumed that $0 \leq k \leq 1$, because in an investigation such as this, ignorance must not be preferred over true belief and error must not be preferred over ignorance. Assuming that the goal is to maximize expected utilities, the values in this matrix determine just one unequivocal minimum confirmation level $L = \max\{k, ½\}$, such that it is rational[2] to

accept H if $\Pr(H) > L$,

reject H if $\Pr(\neg H) > L$

and otherwise to suspend judgment.

How sure we need to be before we accept H is exactly the same as how sure we need to be before we reject H.

The undetermined value of k might seem to be a weak point of the decision matrix. Levi (1962, pp. 56-57) interprets k as the "degree of caution" employed by a researcher. Note that k will realistically be larger than ½, as $k \leq ½$ would mean to automatically recommend accepting a hypothesis as soon as it appears even slightly more likely than its negation. With a view to the often immense consequences of accepting a hypothesis for the continuing search for further truths, it is safe to assume that realistic values of k will be *much* larger than ½, even taking into account only the purely epistemic motives presupposed in our model situation. Levi grants that the choice of k implies that a scientist has to reckon with the seriousness of mistakes, but "only in the sense that the degree of caution that he adopts reflects how serious he considers making *any* mistakes to be in the relation to remaining in doubt" (*ibid*., p. 57). What is particularly relevant in our context is that whatever the choice of k (and however we conceive of the values that must undoubtedly play a role in determining it), a definite level $(1 - L)$ of theoretically *acceptable probability of error* is thereby established and what is more, it is the same level for both types of error. A trade-off between consumer's and producer's risk would not be consistent with a test procedure with purely epistemic aims as characterized by Levi's matrix. So far, it seems that the idea from the beginning of this section could work out: While different levels of acceptable inductive risk are compatible with an investigation solely designed to replace ignorance with true belief and avoid error, the balance of the two types of inductive risk is fixed in such a situation—the acceptable probability of a mistake

[2] I shall here and henceforth exclude from discussion equivocal cases where expected utilities are exactly the same for two different options, as they are of no particular interest for the issues under discussion.

must be the same for both types of error. Any deviation from this balance must from this point of view be seen as reflecting the intrusion of other values in addition to the single epistemic value that this model admits and can in that sense be regarded as bias. I will call this concept of bias "L-bias".

5. Relaxed purity

However, I submit that regarding L-bias as an analysis of real-world cases of preference bias would be much too simple. It presupposes a sense of purity of epistemic activity that is exaggerated and unrealistic. To begin with, it has long been recognized that science, even if conceived as essentially a truth-seeking enterprise, does not pursue each truth with the same eagerness. In the terminology introduced by Philip Kitcher, science aims to find *significant* truths, where significance is bestowed both by a proposition's value for the systematic organization of our beliefs and by our interest in its application (Kitcher 1993, ch. 4, 2001, ch. 6). It is now commonly acknowledged that even basic science must be understood to value a range of properties over and above mere truth: simplicity, unifying power and fruitfulness for the further development of science are examples of such widely accepted "epistemic values" (Kuhn 1977, McMullin 1983). There is no reason to assume that H and $\neg H$ will normally score equally high with regard to all epistemic values, and hence the assumption that both "correct" outcomes must be assigned identical utilities seems amiss even by plausible standards for "pure science".[3] Additionally, a scientist's decision to accept or reject H does not only have consequences for this one particular piece of knowledge but will typically bear on the subsequent development of a research program. Different decisions can open up or foreclose different opportunities to discover more truths. This even holds for the different possible mistakes. Hence, even with regard to replacing ignorance with true belief alone, differences in utility can arise even between the different kinds of mistakes and even between suspending judgment in case H is true and suspending judgment when H is false. Taking all this into account and indicating the utilities of the correct outcomes by c and c', the utilities of the errors by e and e', and the utilities of suspending judgment in different world states by k and k' (all assumed to be represented on an interval scale), do we have anything left to say about the following decision situation that would reflect a realistic ideal of a proper epistemic or scientific enterprise?

[3] Maher (1993, 214-216) tries to defend the assumption by claiming that simplicity is only an instrumental goal of science, pursued solely in virtue of its presumed truth-conduciveness. I cannot imagine how this argument would carry over to other epistemic values like fruitfulness (and find it difficult to defend even for simplicity).

	H is in fact true	H is in fact false
accept H	c	e'
reject H	e	c'
suspend judgment	k	k'

I think we can still plausibly maintain a relaxed ideal of purity by making the following assumptions: The values reflected in the design of a proper scientific investigation should be such that if H is in fact true, accepting it is preferable to suspending judgment, and that is in turn preferable to rejecting H, while if H is in fact false, rejecting H is preferable to suspending judgment, which in turn is preferable to accepting H. That is to say, we can plausibly impose the following order relations on our utilities: $c > k > e$ and $c' > k' > e'$. Under the conditions implied by these, we can formulate a result that follows immediately from the comparison of the expected utilities: An ideal Bayesian agent would

accept H if $\Pr(H) > M_1$,

reject H if $\Pr(\neg H) > M_2$

and otherwise suspend judgment,

where $M_1 = \max\left\{\frac{1}{1+\frac{c-k}{k'-e'}}, \frac{1}{1+\frac{c-e}{c'-e'}}\right\}$ and $M_2 = \max\left\{\frac{1}{1+\frac{c'-k'}{k-e}}, \frac{1}{1+\frac{c'-e'}{c-e}}\right\}$.

To act in accordance with a set of utilities ordered as indicated above, scientists should ideally design their investigation in such a way that it emulates the behavior of this Bayesian agent.

This result shows that a utility structure like the one given above does in fact provide an answer to Rudner's question of how sure we need to be before we accept H in form of a minimum degree of confirmation level of M_1.[4] The exact levels of acceptable error probabilities for the two types of error are given by $(1 - M_1)$ and $(1 - M_2)$ respectively. We see that the balance of inductive risks commanded by our utility structure need not consist in an identity of levels for both types of error probabilities—even if the utilities are interpreted as expressing a purely epistemic interest in the investigation. We can read off how the balance

[4] Note that analogously to the situation in Levi's original matrix, the values for k and k' will in realistic cases be much closer to c and c' respectively than to e and e' respectively, as other constellations would mean a tendency to jump to conclusions on the slightest touch of evidence. Assuming this and a relative similarity between the intervals $c - e$ and $c' - e'$ means that in realistic situations, $M_1 = \frac{1}{1+\frac{c-k}{k'-e'}}$ and $M_2 = \frac{1}{1+\frac{c'-k'}{k-e}}$.

(or rather, imbalance) of risks may legitimately vary with the intervals between c, k and e on the one hand and c', k' and e' on the other.

Does this result permit us to regard the values M_1 and M_2 as indicating the *correct* arrangement of inductive risks, and to consider any infringement thereof a case of preference bias? It seems that the only way to do this would be to claim that for every kind of empirical investigation, there exists an objectively determined set of purely epistemic utilities c, c', e, e', k and k' of its possible outcomes. A deviation of the effective thresholds for acceptance and rejection from M_1 and M_2 would then either be irrational or reflect the intrusion of other, "impure" utilities (subjective "preferences") into the decision problem.

However, the protracted debate on science and values has shown that it is deeply problematic to try and separate epistemic from non-epistemic, or cognitive from non-cognitive values. The common way to attempt such a separation is to claim that epistemic values are those that we believe to indicate the truth of the hypotheses or theories that possess them (McMullin 1983, p. 18). But the truth-conduciveness of simplicity, fruitfulness and the like can hardly be called well established, so "our" belief in their usability as truth-indicator may very well differ from person to person (cf. Kourany 2003, p. 9). For example, feminist scientists and philosophers have criticized the widely shared "epistemic value" of external consistency (of new theories and hypotheses with older, established ones) as being an expression of the value-holders' contentment with the status quo rather than an indicator of truth (cf. Longino 1996, pp. 51-52).

Even if there was a definite set of epistemic *values*, it would still be implausible to assume that this set would objectively determine purely epistemic *utilities* for each and every possible outcome of any investigation. For such epistemic values as we can make out are individually imprecise and can conflict with each other, as has been observed many times (e.g., Kuhn 1977, p. 322, Laudan 1984, pp. 37-38). (Note that cases of conflicting epistemic values can not be shrugged off as freak incidents. Some epistemic values, such as accuracy and breadth of scope, are in systematic tension with each other [cf. Longino 1996, p. 44].) Their application to each individual case must therefore be effected by means of individual judgment.

It may still seem that for all *practical* purposes, we can distinguish between acceptable and unacceptable value influences well enough. It is, after all, exactly the problem of conflicts of interests and bias that research ethicists like David Resnik have in mind when they articulate ethical and epistemological norms to guide and delineate acceptable scientific research (Resnik 2007, esp. ch. 2). Principles such as those proposed by Resnik will certainly find wide support and

can be used to rule out many important cases of objectionable conduct. However, they cannot solve all the problems with the examples introduced in this paper. For example, Doll explicitly justifies his move to exclude certain brain cancer cases from his calculations by referring to the methodological quality of the study that reported them and to the (alleged) fact that this study was itself the origin of the hypothesis to be tested. He thus implicitly appeals to principles of testability and empirical support (to use Resnik's terms, cf. *ibid*., p. 48). So do Wong and Whorton in their remarkable recantation when they claim to be wary of diagnostic bias in their data. The various accepted principles of scientific research leave a lot of leeway for individual decisions to apply or not to apply a certain norm in a certain situation (they *must* leave this leeway, as they often pull in different directions). When this leeway is consistently used in favor of a certain kind of outcome, bias results. The difficulties I have been discussing in this section are thus not merely philosophical, and the efforts to distinguish between epistemic and non-epistemic values or between acceptable and unacceptable principles cannot solve all the problems—neither in theory, nor in practice.

Together, the above considerations show that one cannot mark out one specific set of utilities for a given empirical investigation as the one that is determined by the objective, purely epistemic aims or values; one cannot even impose some restrictions on how such a utility structure is permitted to look (apart from the half order relations we have already been presupposing). Failing any such additional principled constraints on the character of the utilities c, c', e, e', k and k', there remains surprisingly little to say about legitimate test procedures from the perspective of individual rationality. What we *can* say is that there must be thresholds (genuinely between 0 and 1) such that when H's degree of confirmation exceeds the threshold M_1 (respectively falls below $1 - M_2$), the hypothesis should be accepted (or respectively rejected). But these thresholds could be very close to 1 (or respectively 0). The cases discussed in section 1 do not seem to violate this restriction by making it *absolutely impossible* that the unwanted hypothesis will be accepted as a result of the investigation.[5]

Our Bayesian considerations of the researcher's individual rationality thus leave us only with a very weak instrument of criticism. Only if the design of the

[5] For instance, all the epidemiological studies on vinyl chloride mentioned above did end up acknowledging a causal connection between exposure to vinyl chloride and angiosarcoma of the liver (where the evidence was so strong that the strategies of re-interpreting the data that were applied to the case of brain cancer could not have possibly worked). Even in the defectively designed studies on bisphenol A, the probability of registering an effect was arguably not zero, as the strain of rat that was used is apparently not totally unresponsive to estrogenic substances (cf. vom Saal and Hughes 2005, p. 929).

investigation absolutely precludes that acceptance (or rejection) of H could ever emerge as its result, can this be diagnosed as indicating an infringement of the set of conditions $c > k > e$, $c' > k' > e'$ and in that sense reflecting an invalidation of the epistemic character of the investigation. However, in realistic cases such as the two just revisited, it will be plausible to assume that the evaluations at work resulted in utilities that lie very close together for c, k and e, while c', k' and e' are spaced widely apart (taking H to be the hypothesis that the respective substance does have adverse health effects), resulting in a much lower tolerance for the risk of falsely accepting H than for the risk of falsely rejecting H. But to call such an uneven spacing of utilities "bias" would mean disqualifying a lot of research that we usually estimate highly and unhesitatingly consider purely epistemic (think of the utility structure that one must presume to underlie most basic research in physics with regard to the hypothesis: "There exists a uniform representation of the laws of nature").

6. Trust and bias: The perspective of social epistemology

From the vantage point of individualist epistemology, informed by the insight that purist strictures of value-free science cannot be generally upheld, it thus still appears that the cases described at the outset simply reflect the variability of scientific procedure under different admissible value judgments. Remarkably, this is not how the biomedical research community seems to regard the matter. Instead, the community employs a variety of social mechanisms in order to set up conventional standards to prevent just the kinds of phenomena that the cases from section 1 represent.

This has recently been prominently visible in the case of publication bias. To prevent (among other things) the practice of making only favorable results publicly known and sweeping unfavorable ones under the carpet, several organizations have started efforts to register all clinical trials at their inception. The registries are intended to be freely accessible and searchable. Most notably, the International Committee of Medical Journal Editors has made registration a condition of consideration for publication in its member journals.[6] In addition, several organizations of the biomedical community have issued recommendations to

[6] ICMJE 2007, p. 22. Registration of all clinical trials is also endorsed by, e.g., the World Association of Medical Editors and the Association of American Medical Colleges (AAMC) (WAME 2007, Korn and Ehringhaus 2006, p. 2).

academic researchers to make sure that research outcomes cannot be suppressed under the terms of the research contracts they sign.[7]

Aspects of experimental design are also frequently targeted by conventional standards proposed and discussed within the biomedical research community. For example, a peer review panel on low dose effects of endocrine disrupting chemicals (which include bisphenol A), organized in 2001 by the US National Toxicology Program at the request of the Environmental Protection Agency, made several recommendations for future study design, including the following: "Because of clear species and strain differences in sensitivity, animal model selection should be based on responsiveness to endocrine active agents of concern (i.e. responsive to positive controls), not on convenience and familiarity." (NTP 2001, p. vii.)[8] This methodological rule would obviously exclude the CD-SD rat from use for bisphenol A studies. In the case of our second example of biased experimental design, substandard comparison of drugs (like in the case of fluconazole), a respective conventional standard of the biomedical community can be found in §29 of the Declaration of Helsinki: "The benefits, risks, burdens and effectiveness of a new method should be tested against those of the best current prophylactic, diagnostic, and therapeutic methods." (WMA 2004, p. 4.)[9]

Questions regarding the interpretation of outcomes may seem least amenable to clearly expressible standards. A development of interest with respect to the repeated unusual interpretations of epidemiological data concerning the association of vinyl chloride with brain cancer is that more and more editors of medical journals now demand the declaration and publication of authors' financial interests (ICMJE 2007, pp. 8-9, WAME n.d.). In contrast to this, the fact that Richard Doll's vinyl chloride review was commissioned and paid for by the chemical industry was only revealed in a lawsuit twelve years later (Sass et al. 2005, p. 810) and additional heavy financial ties with the chemical industry came out only after Doll's death (Boseley 2006). But even on a more specifically

---

[7] These include the American College of Physicians–American Society of Internal Medicine (Coyle 2002, p. 400) and the AAMC (Korn and Ehringhaus 2006, p. 2). The ICMJE (2007, p. 9) suggests that editors "may choose not to consider an article if a sponsor has asserted control over the authors' right to publish."

[8] This very specific recommendation echoes the common and more general methodological advice to take special care in choosing proper controls in animal research, and to take into account that "[m]ore than one control is frequently required" (Johnson and Besselsen 2002, p. 206). Cf. also Festing and Altman 2002, p. 248.

[9] While this principle (sometimes called "equipoise" or the "uncertainty principle") was originally introduced as an *ethical* principle in order to protect trial participants, its *methodological* relevance is now widely recognized in the biomedical research community (cf. Dietz 2007, Djulbegovic et al. 2000).

methodological level, seminal attempts to propose standards for preventing bias can be found. In a methodological paper discussing review articles and meta-analyses, an international group of epidemiologists suggests that each epidemiological overview needs an explicit description in the study protocol of how studies are selected to be included in the overview and that the analysis should then make use of "all studies that are relevant according to the explicit inclusion criteria" (Blettner et al. 1999, p. 3). The consistent adherence to this admonition would prevent the kind of ad hoc re-interpretation of the database for a single result which allowed Doll to downplay mortality from brain cancer in his vinyl chloride review.

It appears that for each of the cases of preference bias we have considered, there exist conventional standards that have been proposed and discussed in the biomedical research community which would, if adhered to, act as countermeasures to these kinds of phenomena. The conventional standards are methodological, but in a broader sense (since they also include recommendations concerning the dissemination and publication of results), and they come in varying degrees of bindingness, explicitness and generality. They do not generally have the status of strict, codified and generally accepted methodological rules; it is plausible that some of them are hardly ever explicitly formulated except when an infringement needs to be criticized.

Nonetheless, the conventional standards are discernible, and the fact that members of the research community appeal to them when criticizing the work of others testifies to their relevance. In the present context, the existence of conventional standards poses two important questions. Firstly, the presence of conventional standards within the respective research community seems to reinforce the intuitive feeling that the examples we have considered which infringe upon such standards constitute instances of epistemic failure. Why? Secondly, the standards all impose implicit restrictions on the ways in which the free exertion of evaluative judgment may determine the balance of inductive risk (as considered in section 5). If value judgments are part and parcel of the research process at virtually each and every step, as has been suggested, then how can such restrictions be justified?

In the remainder of this paper, I would like to propose an analysis of conventional standards of a research community that will turn them into a key element for understanding the epistemological failure in preference bias and at the same time will serve to answer the two questions of the preceding paragraph. This analysis will require us to give up the perspective of individual rationality and consider the wider picture of social epistemology instead. I suggest that a

conventional standard represents an effort of a research community's members to coordinate their practices in order to enable and preserve epistemic trust in their research results.

Trust plays a decisive role for the social epistemology of the sciences for at least two (interrelated) reasons: First of all, the division of cognitive labor within science requires researchers to estimate reliably the dependability of each other's results. In addition, science has epistemological roles for society at large—specific roles that are central to the knowledge situation of each member of society, similar in kind (if not in the specifics) to the special roles of journalism. Roughly speaking, these roles center on the production of a certain kind of new knowledge with a high degree of reliability. One need not spell this out more precisely in order to realize that the fulfillment of such roles requires widespread epistemic trust in the results of scientific research. As a matter of fact, the importance of trust is often emphasized in the discussion of conventional standards. For instance, in a background report to one of the more explicitly normative documents we have been drawing from above (see notes 6 and 7), the authors, who represent a task force of the Association of American Medical Colleges, declare as their objective "to preserve public trust in clinical research while sustaining medical progress" (AAMC 2001. p. 1). They explain that "the public insists that universities […] continue to serve society as trusted and impartial arbiters of knowledge" (*ibid*., p. 24).

Epistemic trust in research outcomes can only develop and stabilize if it is possible to make *realistic* estimates of their dependability. The social epistemology of science thus requires that all kinds of actors within and without science develop differentiated attitudes of confidence towards different kinds of institutionally sanctioned scientific "results". But typically they cannot appraise the underlying value-judgments at work. This threatens to thwart any attempt to assess the dependability of a result, because as we have seen, value structures have a strong bearing on the admissible error probabilities within a test procedure. In this situation, research communities adopt conventional standards that impose implicit constraints on acceptable error probabilities. Ideally this makes it possible for individual epistemic actors to develop a reliable sense for the dependability of certain kinds of scientific outcomes (such as the conclusions of epidemiological meta-analyses) on the basis of their experience and their knowledge of the procedures, but without knowing in each case which of the outcomes would have been the one preferred by the investigators. The conventional standards are solutions to coordination problems (as, according to David Lewis [1969, ch. 1], all conventions are). It does not matter so much which specific balance of inductive

risks is implied by the standard, but rather that there *is* a standard which is dependably observed by everyone in the community. With the help of such entrenched standards, users can learn to appraise the reliability of different kinds of research results. Obviously, this still requires a lot of skill and knowledge, as standards can differ for different kinds of procedures and different research communities. But the crucial advantage remains that users do not have to guess at the researchers' value-judgments. The standards adopted are arbitrary in the sense that there could have been a different solution to the same coordination problem, but once a specific solution is socially adopted, it is in a certain sense binding. For example, how grave an epistemic mistake it is to ignore positive controls is determined by an epistemic environment in which consideration of all relevant controls is taken for granted (and therefore relied upon)—just as how grave an endangerment of traffic it is to drive on the left is relative to a social environment that strictly obeys the rule to drive on the right. I am not arguing that all conventional methodological standards that are actually in effect are *optimal* solutions to the problem of coordinating the collective effort to enable and preserve epistemic trust—research communities obviously sometimes amend and change standards. But as long as a conventional methodological standard is well-entrenched and not openly challenged, that fact itself acquires a specific social-epistemological significance. It is the disregard of important and entrenched collective trust-enabling measures that we perceive as epistemic failure in cases where the conventions are disobeyed.

It might be tempting to speculate that the conventional standards should typically strike a symmetric balance of inductive risks, i.e. that they should tend to avoid L-bias and thereby employ epistemic purity (in the sense discussed in section 4) as a regulative ideal.[10] In practice, however, many standardized research practices can not be regarded in that way. For example, significance testing, the most common statistical approach, is designed to provide "severe tests" only in one sense, namely that they impose a demanding evidential hurdle for *accepting* the claim that an effect actually exists. On the other hand, such tests normally are unable to provide a well-defined distinction between *suspending judgment* and *rejecting* the claim that there is an effect. The method is clearly designed to avoid the mistake of falsely admitting the existence of an effect where really there is none (type I error) and implicitly presupposes that this is the more serious one of the two errors (cf. Levi 1962, pp. 58-63). As they stand, classical significance tests are

[10] This might often be the proclaimed aim of documents proposing conventional standards, which frequently contain the assertion that the problem at hand was *bias* and how to avoid it. Cf. ICMJE 2007, WAME n.d., Korn and Ehringhaus 2006, p. 1, AAMC 2001, p. 1, Coyle 2002, p. 400.

therefore unsuitable for the avoidance of L-bias. This is notable, because significance testing is itself an important standardized procedure and many conventional standards of hypothesis testing are embedded in its methodology, e.g. the widespread conventional choice of .05 as the highest reasonable significance level. Clearly, such conventional standards as those related to significance testing also place constraints on the implicit utility structure of the test situation and on the admissible error probabilities. Its implicit preference structure is not symmetric, but nevertheless its specific asymmetry is one that is well-entrenched in scientific practice and is thereby trust-enabling.[11]

As we have seen, each of the cases considered in section 1 breaches one or the other convention of the kind that is adopted by research communities in order to be able to fulfill their social epistemological roles. This insight can finally lead to a more refined definition of preference bias: Preference bias is the infringement of an explicit or implicit conventional standard of the respective research community in order to increase the likelihood of arriving at a preferred result. The intuition that preference bias constitutes epistemic failure (rather than just being a matter of differing value judgments) can thus finally be captured within the framework of social epistemology. Note that the fact that the conventional standards are sometimes vague and often only implicit accounts for a wide grey zone in the identification of preference bias.

Preference bias, thus defined, shows that *cases exist* in which the shifting of the balance of inductive risks in accord with individual preferences is rightfully regarded as an unequivocal epistemic mistake, regardless of the values we happen to hold. I do not claim that this concept of preference bias covers *all* possible epistemological problems caused by undue influence of preferences. For instance, it might be claimed that sometimes the conventional standards of a research community are themselves distorted by interests and preferences in an epistemologically problematic way.[12] Obviously, this kind of claim will have to await an explication that differs from my analysis of preference bias.

Note that I am not at all arguing that science is free of individual value-judgment after all, or even that the ways in which value-judgments can inform the balance of inductive risks are always conventionally restricted. I do not believe that crucial decision situations in science are generally amenable to rules and standards

---

[11] I am indebted to Birgitte Wandall for helpful discussion on an earlier, misguided predecessor of this paragraph.

[12] This possibility might be illustrated by Shrader-Frechette's (2004) claim that many shortcomings of the suggestions of the International Commission of Radiological Protection can be traced back to problematic methodological assumptions employed in the research underlying them.

in a similar way as more or less standardized procedures. I do not think there can be value-free decisions, for example, to adopt a whole theory, or to accept or reject a novel method. The acceptable error probabilities in these areas of scientific practice are at best constrained by the conditions of relaxed purity discussed in section 5. Nevertheless, there are many procedures and aspects of research that are subject to stronger constraints, imposed by entrenched conventional standards of the community.

One point that remains to be addressed is the possible concern that this analysis of preference bias is somehow not social enough for an analysis from the vantage point of social epistemology. After all, it builds on conventional standards which, while socially constituted, could in principle be observed by individuals. In contrast, other social epistemologists have maintained that science asserts or should assert its objectivity through norms and procedures that themselves operate on a social level, by guaranteeing such social principles as equality of intellectual authority and recognized avenues of criticism within a pluralistic scientific community (Longino 1990, ch. 4).[13] But surely, the research community can and will take *diverse* measures in order to preserve the trust that is the basis of their existence. These will plausibly include both methodological standards *and* norms that affect social organization and interaction.[14] In this paper, my concern was to find out what is wrong with a certain kind of phenomenon that bears the appearance of epistemic failure—even if taken individually. (One does not seem to need to consider the question of how many other non-industry studies on bisphenol A there were in order to find fault with the industry-funded ones.) My claim is that *in these cases*, the problem lies with the violation of conventional methodological standards and that the latter should be regarded as efforts to coordinate admissible error probabilities for certain procedures within a community. The analysis is therefore an essentially social one.

The parallel existence of "methodological" and other, more obviously "social" safeguards of trust and objectivity is also good to bear in mind when it comes to the consideration of practical consequences. The fact that the original sin in cases of preference bias is the infringement of a conventional methodological standard of the research community does not imply that the most effective remedy against the widespread phenomenon of preference bias may not operate on the social level. Such social measures could for example consist in efforts to

[13] Similarly, Miriam Solomon (2001, esp. ch. 8) finds no fault with individual biases (except that she objects to the term "bias") and even regards them as productive as long as they are not poorly distributed within the community.

[14] On Longino's own view (2002, pp. 130-131, 145-146), shared public standards of inquiry play an essential role in shaping and defining a cognitive community.

counterbalance the increasing amount of research funded by interested parties with publicly funded research, or otherwise to institutionally organize the carrying out of, say, clinical trials in a way that precludes sponsors' influence on the research outcomes (cf. Biddle 2007).

7. Conclusions

I have maintained that preference bias consists in the infringement of conventional standards entertained by the respective research community. This analysis captures the intuition that preference bias constitutes an epistemic shortcoming, as the conventional standards themselves are adopted by the community in an effort to make possible and preserve epistemic trust and to ensure the community's capability of fulfilling its epistemological roles. It also explains why the diagnosis of preference bias is often not a clear-cut case, as the conventional standards at issue come in different degrees of explicitness and universality.

As a second conclusion, we should note that an analysis of preference bias as an epistemic shortcoming was only possible by viewing it from the perspective of social epistemology. The differing frameworks of individual rationality we considered were instructive with regard to the connection between inductive risk and certain concepts of bias, but they did not offer us any definitive and realistic constraints in order to draw a line between the inevitable value-ladenness of science and unacceptable preference bias. According to the picture that has emerged from our investigation, that line is only socially constituted through the development of conventional trust-preserving standards.

The third and last conclusion which I would like to stress is that these standards reveal continuing restrictions of the epistemologically legitimate control of researchers' individual value-judgments over the balance of inductive risks. The dominion of the standards is limited to certain procedures and aspects of the research process that are particularly amenable to regulation by explicit and implicit rules. But as the examples discussed in this paper show, these limited aspects can sometimes be of vital importance. Though the criticism of the traditional conception of value-free science has brought important insight, a picture of science as an open playing-field for individual value-judgments would therefore be exaggerated.

Acknowledgements

I would like to thank Justin Biddle, Jim Brown, Martin Carrier, Cornelis Menke, Birgitte Wandall, Ken Westphal, Eric Winsberg, Alison Wylie and an anonymous referee for this journal for their helpful remarks on earlier versions of this paper.

References

AAMC (2001). *Protecting subjects, preserving trust, promoting progress–Policy and guidelines for the oversight of individual financial interests in human subjects research*. Association of American Medical Colleges, AAMC Task Force on Financial Conflicts of Interest in Clinical Research. http://www.aamc.org/research/coi/firstreport.pdf. (Accessed 11 March 2008)

Barrow, C. S. & Conrad, J. W. (2006). Assessing the reliability and credibility of industry science and scientists. *Environmental Health Perspectives*, *114* (2), 153-155.

Bekelman, J. E., Li, Y. & Gross, C.P. (2003). Scope and impact of financial conflicts of interest in biomedical research: A systematic review. *Journal of the American Medical Association*, *289*, 454-465.

Biddle, J. (2007). Lessons from the Vioxx debacle: What the privatization of science can teach us about social epistemology. *Social Epistemology*, *21* (1), 21-39.

Blettner, M., Sauerbrei, W., Schlehofer, B., Scheuchenpflug, T., & Friedenreich, C. (1999). Traditional reviews, meta-analyses and pooled analyses in epidemiology. *International Journal of Epidemiology*, *28*, 1-9.

Blumenthal, D., Campbell, E. G., Anderson, M. S., Causino, N. & Louis, K. S. (1997). Withholding research results in academic life science: Evidence from a national survey of faculty. *Journal of the American Medical Association*, *277*, 1224–1228.

Boseley, S. (2006). Renowned cancer scientist was paid by chemical firm for 20 years. *The Guardian*, 8 December 2006, p. 1.

Chan, A.-W., Hróbjartsson, A., Haahr, M. T., Gøtzsche, P. C. & Altman, D. G. (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. W. (1948). *Theory of experimental inference*. New York: Macmillan.

Coyle, S. L. (2002). Physician–industry relations. Part 1: Individual physicians. *Annals of Internal Medicine*, *136* (5), 396-402.

Dietz, H. P. (2007). Bias in research and conflict of interest: Why should we care? *International Urogynecology Journal*, *18*, 241-243.

Djulbegovic, B., Lacevic, M., Cantor, A., Fields, K. K., Bennett, C. L., Adams, J. R., Kuderer, N. M. & Lyman, G. H. (2000). The uncertainty principle and industry-sponsored research. *The Lancet*, *356*, 635-638.

Doll, R. (1988). Effects of exposure to vinyl chloride: An assessment of the evidence. *Scandinavian Journal of Work, Environment & Health*, *14* (2), 61-78.

Douglas, H. (2000). Inductive risk and values in science. *Philosophy of Science*, *67*, 559-579.

Festing, M. F. W. & Altman, D. G. (2002). Guidelines for the design and statistical analysis of experiments using laboratory animals. *ILAR Journal*, *43* (4), 244-258.

Gluud, L. L. (2006). Bias in clinical intervention research. *American Journal of Epidemiology*, *163* (6), 493-501.

Goldman, A. I. (1999). *Knowledge in a social world*. Oxford: Oxford University Press.

Hempel, C. G. (1965). Science and human values. In *Aspects of scientific explanation* (pp. 81-96). New York: Free Press.

ICMJE (2007). *Uniform requirements for manuscripts submitted to biomedical journals: Writing and editing for biomedical publication*. Updated October 2007. International Committee of Medical Journal Editors. http://www.icmje.org/icmje.pdf. (Accessed 11 March 2008)

Jeffrey, R. C. (1956). Valuation and acceptance of scientific hypotheses. *Philosophy of Science*, *23* (3), 237-246.

Johansen, H. K. & Gøtzsche, P. C. (1999). Problems in the design and reporting of trials of antifungal agents encountered during meta-analysis. *Journal of the American Medical Association*, *282*, 1752-1759.

Johnson, P. D. & Besselsen, D. G. (2002). Practical aspects of experimental design in animal research. *ILAR Journal*, *43* (4), 203-206.

Kitcher, P. (1993). *The advancement of science*. Oxford: Oxford University Press.

Kitcher, P. (2001). *Science, truth and democracy*. Oxford: Oxford University Press.

Kjaergard, L. L. & Als-Nielsen, B. (2002). Association between competing interests and authors' conclusions: Epidemiological study of randomized clinical trials published in the BMJ. *British Medical Journal*, *325*, 249-252.

Korn, D. & Ehringhaus, S. (2006). Principles for strengthening the integrity of clinical research. *PLoS Clinical Trials*, 1 (1), e1. doi:10.1371/journal.pctr.0010001.

Kourany, J. A. (2003). A philosophy of science for the twenty-first century. *Philosophy of Science*, *70* (1), 2003, 1-14.

Krimsky, S. (2003). *Science in the private interest: Has the lure of profits corrupted biomedical research?* Lanham, MD: Rowman & Littlefield.

Kuhn, T. S. (1977). Objectivity, value judgment, and theory choice. In *The essential tension* (pp. 320-339). Chicago: University of Chicago Press.

Laudan, L. (1984). *Science and values*. Berkeley: University of California Press.

Levi, I. (1960). Must the scientist make value judgments? *The Journal of Philosophy*, *57* (11), 345-357.

Levi, I. (1961). Decision theory and confirmation. *The Journal of Philosophy*, *58* (21), 614-625.

Levi, I. (1962). On the seriousness of mistakes. *Philosophy of Science*, *29* (1), 47-65.

Lewis, D. K. (1969). *Convention: A philosophical study*, Cambridge, MA: Harvard University Press.

Lexchin, J., Bero, L.A., Djulbegovic, B. & Clark, O. (2003). Pharmaceutical industry sponsorship and research outcome and quality: Systematic review. *British Medical Journal*, *326*, 1167-1170.

Longino H. E. (1990). *Science as social knowledge: Values and objectivity in scientific inquiry*. Princeton: Princeton University Press.

Longino, H. E. (1996). Cognitive and non-cognitive values in science: Rethinking the dichotomy. In L. Henkinson Nelson & J. Nelson (Eds.), *Feminism, science, and the philosophy of science* (pp. 39-28). Dordrecht: Kluwer.

Longino, H. E. (2002). *The fate of knowledge*. Princeton: Princeton University Press.

Maher, P. (1993). *Betting on theories*. Cambridge: Cambridge University Press.

Markowitz, G. & Rosner, D. (2002). *Deceit and denial: The deadly politics of industrial pollution*. Berkeley: University of California Press.

McMullin, Ernan (1983). Values in science. In P. D. Asquith & T. Nickles (Eds.), *PSA 1982*, vol. 2 (pp. 3-28). East Lansing: Philosophy of Science Association.

Mundt, K. A., Dell, L. D., Austin, R. P., Luippold, R. S., Noess, R. & Bigelow, C. (2000). Historical cohort study of 10 109 men in the North American vinyl chloride industry, 1942–72: Update of cancer mortality to 31 December 1995. *Occupational and Environmental Medicine*, *57* (11), 774-781.

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-server.niehs.nih.gov/ntp/htdocs/liason/LowDosePeerFinalRpt.pdf. (Accessed 11 March 2008)

Pearson, E. S. (1933). A survey of the uses of statistical method in the control and standardization of the quality of manufactured products. *Journal of the Royal Statistical Society*, *96* (1), 21-75.

Resnik, D. (2000). Financial interest and research bias. *Perspectives on Science*, *8* (3), 255-285.

Resnik, D. (2007). *The price of truth: How money affects the norms of science.* Oxford: Oxford University Press 2007.

Rudner, R. (1953). The scientist *qua* scientist makes value judgments. *Philosophy of Science*, *20* (1), 1-6.

Sass, J. B., Castleman, B. & Wallinga, D. (2005). Vinyl chloride: A case study of data suppression and misrepresentation. *Environmental Health Perspectives*, 113, 809-812.

Shrader-Frechette, K. (2004). Using metascience to improve dose-response curves in biology: Better policy through better science. *Philosophy of Science*, *71* (5), 1026-1037.

Solomon, M. (2001). *Social empiricism*. Cambridge, MA: MIT Press.

Song, F., Eastwood, A. J., Gilbody, S., Dulea, L. & Sutton, A.J. (2000). Publication and related biases. *Health Technology Assessment*, *4* (10), 1-115.

Vom Saal, F. S. & Hughes, C. (2005). An extensive new literature concerning low-dose effects of bisphenol A shows the need for a new risk assessment. *Environmental Health Perspectives*, *113*, 926-933.

Vom Saal, F. S. & Welshons, W. V. (2005). Large effects from small exposures. II. The importance of positive controls in low-dose research on bisphenol A. *Environmental Research*, 100 (1), 50-76.

Wald, A. (1942). *On the principles of statistical inference* (= *Notre Dame Mathematical Lectures* 1). Notre Dame, IN: University of Notre Dame.

WAME (2007). *WAME policy statements, prepared by the Editorial Policy Committee*. World Association of Medical Editors. http://www.wame.org/resources/policies. (Accessed 11 March 2008)

WAME (n.d.). *WAME publication ethics policies for medical journals*. World Association of Medical Editors. http://www.wame.org/resources/ethics-resources/publication-ethics-policies-for-medical-journals. (Accessed 11 March 2008)

Wong, O., Whorton, M. D., Foliart, D. E. & Ragland, D. (1991). An industry-wide epidemiologic study of vinyl chloride workers, 1942–1982. *American Journal of Industrial Medicine*, *20* (3), 317-334.

Wong, O. & Whorton, M. D. (1993). Diagnostic bias in occupational epidemiologic studies: An example based on the vinyl chloride literature, *American Journal of Industrial Medicine*, *24* (2), 251-256.

WMA (2004). *World Medical Association declaration of Helsinki*. WMA General Assembly. http://www.wma.net/e/policy/pdf/17c.pdf. (Accessed 11 March 2008)