

How strong is the argument from inductive risk?

Abstract: The argument from inductive risk, as developed by Rudner and others, famously concludes that the scientist qua scientist makes value judgments. The paper aims to show that trust in the soundness of the argument is overrated – that philosophers who endorse its conclusion (especially Douglas and Wilholt) fail to refute two of the most important objections that have been raised to its soundness: Jeffrey's objection that the genuine task of the scientist is to assign probabilities to (and not to accept or reject) hypotheses, and Levi's objection that the argument is ambiguous about decisions about how to act and decisions about what to believe, that only the former presuppose value judgments, and that qua scientist, the scientist only needs to decide what to believe.

1. Introduction

It has become customary to divide the process of scientific discovery into three stages or contexts. In the context of discovery, objects of investigation are selected in accordance with considerations of relevance; in the context of justification, data are gathered to confirm or disconfirm hypotheses; and in the context of application, hypotheses are applied for various social or political purposes. It is widely agreed that value judgments operate in the contexts of discovery and application. Judgments about the respective value of objects of investigation at least partly determine the selection of these objects, and judgments about the values of social or political purposes play an important role when it comes to applications of scientific hypotheses. It is also widely agreed that value judgments *necessarily* operate in the contexts of discovery and application: that questions of relevance cannot be decided independently of value judgments, and that applications of scientific hypotheses presuppose value judgments. What is less widely agreed is whether value judgments necessarily operate in the core of science, i.e. the context of justification. Many theorists believe that value judgments operate in the context of justification as a matter of fact. But what remains contested is whether value judgments *necessarily* operate in the context of justification: whether scientists are *incapable* of confirming or disconfirming hypotheses independently of the value judgments that they happen to endorse.

The so-called argument from inductive risk (AIR) has been designed to support precisely this claim. It concludes that value judgments necessarily operate even in the core of science, that scientists necessarily make value judgments, or that the scientist qua scientist makes value judgments (and not qua the kind of person that the scientist happens to be). AIR may thus be read as a direct attack on the value-free ideal, i.e. an ideal that can be understood as the

requirement that a specific realm of scientific activity, namely the context of justification, be free from science-external value judgments (cf. e.g. Schurz, 2014, p. 42).

In the context of AIR, confirmation is understood as the assignment of a probability that is decided to be sufficiently high to warrant the acceptance of a scientific hypothesis. Variants of AIR can be found in Churchman (1948) and Hempel (1965). But the variant that is typically discussed in more recent contributions to the debate goes back to a six-page essay by Rudner (1953). The variant says that

- (1) any scientist S rejects or accepts hypotheses *qua* scientist,
- (2) S accepts (rejects) hypothesis H iff S can assign a probability P to H and decides that P is (not) sufficiently high to warrant the acceptance of H ,
- (3) S 's decision whether P is (not) sufficiently high presupposes value judgments,
- (4) therefore, S makes value judgments *qua* scientist.

Hempel (1965, p. 92) is the one who speaks of the “inductive risk” of accepting (rejecting) false (true) hypotheses. Accordingly, the derivation of (4) from (1) – (3) has become known as “argument from inductive risk”.

At first sight, AIR appears to be quite strong. It seems that the acceptance or rejection of hypotheses and the assignment of probabilities are among the most genuine activities of scientists. Another strength is that the assignment of probabilities can be understood as inductive inference in the broadest sense of the term. The argument seems to remain sound, that is, if “probability” is replaced with “degree of confirmation” or “degree of belief” in premises (2) and (3). It also seems that the acceptance or rejection of hypotheses depends on decisions about the sufficiency of probabilities, and that these decisions depend on value judgments.

That chain of dependencies can be illustrated in various ways. Consider, for instance, the hypothesis H that a sample of a polio vaccine is free from active polio virus (cf. Jeffrey, 1956, pp. 241-2). If H were to be tested on a treatment group consisting entirely of mice, S would perhaps decide that a probability of 0.95 is sufficiently high to warrant the acceptance of H . If, however, H were to be tested on a treatment group consisting entirely of children, she would most probably decide that a probability of 0.95 is not sufficiently high. Whether or not she accepts H , depends on her decision about the sufficiency of P , and her decision about the sufficiency of P depends on a value judgment about the seriousness of acting on the basis of H if H is false.

At the same time, the conclusion of the argument appears to be quite radical. Philosophers who have defended that conclusion more recently somewhat downplay its radicalness when arguing that accepting “the role of values in science does not eliminate the requirement for

good arguments” (Douglas, 2000, p. 600), or that the “value judgments pervading scientific practice” are “reasonable” (Kitcher, 2011, p. 36). But the radicalness of the conclusion is adequately expressed by the remark by which Rudner (1953, p. 6) closes his essay: “[I]f the major point I have here undertaken to establish is correct, then clearly we are confronted with a first order crisis in science & methodology.” What the participants in the debate call “values” or “value judgments” are non-scientific values that include ideologies and material interests after all.

The radicalness of the conclusion explains why some philosophers felt early on that they needed to object to the validity or soundness of AIR. Jeffrey (1956, p. 237), for instance, objects to premise (1) that “the activity proper to the scientist is the assignment of probabilities [...] to [...] hypotheses”, and not the acceptance or rejection of hypotheses. Levi (1962, p. 47), by contrast, can be read as objecting to a purported ambiguity in (2) and (3): as objecting that the decision referred to in (2) is a decision about what to believe, while the decision referred to in (3) is a decision about how to act, that only the latter presupposes value judgments, and that the scientist qua scientist only needs to decide what to believe. It is worth mentioning that Hempel (1949, p. 560), who is sometimes viewed as a proponent of AIR (cf. e.g. Douglas, 2000, pp. 560-2), raises an objection that is similar to Jeffrey’s in his review of Churchman’s *Theory of Experimental Inference*.¹

But most philosophers participating in the debate believe that these objections can be refuted. Rudner (1953, p. 4) anticipates Jeffrey’s objection when suggesting that assigning probability P to hypothesis H is the same as accepting the hypothesis H_1 that the probability of H is P , and that accepting H_1 presupposes value judgments. Jeffrey (1956, p. 246) responds that assigning probability P to H is *not* the same as accepting H_1 . But Douglas (2009, 53-4) thinks that Jeffrey’s response leads into a vicious regress of assigning probabilities to probability assignments. Douglas (2016, pp. 614-5) and Wilholt (2009, p. 94) argue against Jeffrey that scientists *should* accept or reject hypotheses because they are responsible for the actions that are taken on the basis of these hypotheses. And Wilholt (2009, pp. 95-6) argues against Levi that his conception of a decision about what to believe “presupposes a sense of purity of epistemic activity that is exaggerated and unrealistic”.

The present paper aims to show that Rudner, Douglas, and Wilholt do not manage to refute the objections raised by Jeffrey and Levi, and that trust in the soundness of AIR is accordingly overrated. The paper will argue, more specifically, that

¹ Carnap (1950, pp. 205-7) happens to be another philosopher endorsing the view that the activity proper to the scientist is to assign probabilities (or “degrees of confirmation”) to hypotheses and not to accept or reject them. Levi (1960, p. 353) therefore likes to speak of the “Carnap-Hempel-Jeffrey view”.

- assigning probability P to H without accepting the hypothesis H_1 that the probability of H is P is unlikely to lead into a vicious regress of assigning probabilities to probability assignments.
- scientists should accept or reject hypotheses only hypothetically, if they should accept or reject hypotheses at all.
- decisions about what to believe might not be epistemically pure in the sense envisaged by Levi, but that what supplements decisions about what to believe may be pragmatic considerations that do not necessarily involve value judgments.

The argument that scientists should accept or reject hypotheses only hypothetically is in fact an argument that Schurz (2013, pp. 325-8) develops in a paper that has appeared only in German, and that he mentions only briefly in his textbook on philosophy of science (cf. Schurz, 2014, p. 77). Schurz (2013, p. 330) also explains why we may concentrate on the inductive risk of accepting (rejecting) false (true) hypotheses when evaluating the soundness of AIR. Douglas (2000, p. 563) is right when observing that scientists “also take inductive risks in stages of science before acceptance or rejection of theories”. But these risks add up to the total risk of the scientific hypothesis that is relevant to public decision-making. And thus, it is not entirely correct to say that these risks are “never brought to the light of public decision-making” (Douglas, 2000, p. 563).

Note that the paper does not exactly aim to explicitly defend Jeffrey’s or Levi’s objection. Defending both objections would be self-defeating because, as Levi (1960, p. 345; 1962, p. 47) points out on several occasions, the view that the activity proper to the scientist is not the acceptance or rejection of hypotheses is incompatible with the view that the scientist qua scientist needs to decide what to believe. But the paper is not going to defend any one of the two objections either. It will remain undecided as to whether the activity proper to the scientist is the assignment of probabilities or the acceptance and rejection of hypotheses.² What the paper aims to show is that AIR is weaker than often suggested because Rudner, Douglas, and Wilholt fail to refute Jeffrey’s and Levi’s objections.

The paper is organized as follows. Section 2 will reproduce Douglas’s discussion of experiments that use laboratory animals to find out whether dioxins cause liver cancer because that discussion nicely illustrates the various stages of hypothesis confirmation in which inductive risk is present. But section 2 will also defend Schurz’s claim that the inductive risk that is present in these stages adds up to the total risk of the scientific hypothesis that is relevant to public decision-making. Section 3 will argue that assigning

² The opposition between the two views appears to be a symptom of another (potentially irresolvable) dispute – that between antirealism and realism. While Jeffrey might have inherited the antirealism of the logical empiricists (most notably Carnap and Hempel), Levi (1984) endorses a realist position called “myopic realism”.

probability P to H without accepting the hypothesis H_1 that the probability of H is P is unlikely to lead into a vicious regress, and that scientists should accept or reject hypotheses only hypothetically, if it is true that qua scientists, they should accept or reject hypotheses. Section 4 will try to show that the Higgs boson discovery exemplifies a decision about what to believe on the part of physicists: a decision that is epistemically pure in the sense of not depending on value judgments, but not epistemically pure in the sense of not depending on pragmatic considerations. Section 5 will argue that decisions about what to believe that are epistemically pure in this sense are also possible in sciences with clear non-epistemic impacts. Section 6 will conclude by briefly summarizing the line of argument and main findings of the paper, while also pointing to a number of limitations and questions for future research.

2. Inductive risk: partial and total

Douglas (2000, p. 563) argues, to repeat, that scientists “also take inductive risks in stages of science before acceptance or rejection of theories”. She distinguishes three such stages: one of selecting an appropriate level of statistical significance, one of evidence characterization and one of the interpretation of results. And she argues that value judgments necessarily operate in all three stages.

Douglas illustrates her distinction by discussing experiments that use laboratory animals to find out whether dioxins cause liver cancer. In experiments of this kind, lab rats are assigned to a treatment and a control group by some randomizing procedure. Rats in the treatment group are dosed or exposed to dioxin, while rats in the control groups remain non-dosed. Since both control and dosed animals usually exhibit some liver cancer, what needs to be determined is whether the cancer rate among the exposed animals is significantly higher than the cancer rate among the control animals. Unless it is significantly higher, it cannot be considered a genuine result of exposition to dioxin.

Douglas doesn't use the term of “null hypothesis” even once. But for reasons that will become clear in section 5, researchers should say that in experiments that use laboratory animals to find out whether dioxins cause liver cancer, the null hypothesis is (or should be) that dioxins don't cause liver cancer. Douglas argues that whenever a hypothesis is tested, two levels of significance need to be determined: level α for the maximum probability of committing a type I error of rejecting the hypothesis if it is true, and level β for the maximum probability of committing a type II error of accepting the hypothesis if it is false. High α will lead to an excess of false positives, and high β to an excess of false negatives.

Since the null hypothesis says that dioxins don't cause liver cancer, an excess of false positives will lead to an over-regulation of the dioxin-producing parts of the industry because dioxins appear more toxic than they actually are. By contrast, an excess of false negatives will lead to an under-regulation of the dioxin-producing parts of the industry because dioxins appear less toxic than they actually are. And while an over-regulation is more in the interest of the general public, an under-regulation is more in the interest of the industry.

There is a tricky issue in statistical hypothesis testing: the two levels of significance α and β trade off each other – the greater α (β), the smaller β (α). And there seems to be no way to determine α and β without passing a value judgment about the respective merits of under- or overregulating the dioxin-producing parts of the industry. According to Douglas, this is the first lower stage at which value judgments operate: whenever we determine the maximum probability of committing a type I (II) error, we need to evaluate the consequences of an excess of false positives (negatives).

The second lower stage, at which Douglas thinks value judgments operate, is that of evidence characterization. In the case of lab animal experiments that test the hypothesis that dioxins don't cause liver cancer, that is the level of expert pathologists judging whether a tissue sample has a cancerous lesion or not. Douglas asks readers to consider two extreme cases. In the first case, the pathologist judges all borderline cases to be non-cancerous lesions. Such an approach leads to an excess of false negatives and an under-regulation of the dioxin-producing parts of the industry. In the second case, the pathologist judges all borderline cases to be malignancies. Such an approach leads to an excess of false positives and an over-regulation of the dioxin-producing parts of the industry.

Douglas argues that what a consideration of these cases tells us is that pathologists are similarly concerned with false positives and false negatives, though not as formally as at the level of selecting significance levels: that whenever judging whether a borderline case has a cancerous lesion or not, pathologists need to evaluate the consequences of an excess of false positives or false negatives. If they prioritize the interests of the industry (the general public), they will tend to identify borderline cases as non-cancerous (malignant).

The third lower stage, at which Douglas thinks value judgments operate, is that of the interpretation of results. In the case of lab animal experiments that test the hypothesis that dioxins don't cause liver cancer, that is the level of answering the question whether there is a threshold for the carcinogenic effects of dioxins, where a threshold is understood as one of dosing the rats at x ng/kg/day. Douglas (2000, pp. 573-4) argues that there are two competing interpretations that give rise to a positive and a negative answer, respectively. The one giving rise to the positive answer relies on the Paracelsian assumption that the right

dose differentiates the poison and the remedy. The one giving rise to the negative answer, by contrast, relies on the empirical finding that there is no threshold for the carcinogenic effects of radiation: great and small amounts of radiation do and do not cause cancer.

Which of the two interpretations should we endorse? As there isn't any fact of the matter that tells us which of them to prefer, we will take our pick in accordance with how we evaluate the consequences of our choices. If we falsely believe that there is a threshold, regulations will probably protect public health insufficiently; and if we falsely believe that there is no threshold, regulations will probably be overly stringent. Therefore, the interpretation we endorse will depend on whether we put the interests of the general public before the interests of the industry, or vice versa. The third lower stage is interesting but can be ignored in the remainder because the question of there being a threshold for the carcinogenic effects of dioxins is specific to the case of lab animal experiments that test hypotheses of the toxicity of substances.

Douglas (2000, p. 565) concludes that "just as there is inductive risk for accepting theories, there is inductive risk for accepting methodologies, data, and interpretations". She moves on to say that by "expanding where we see relevant inductive risk, the potential role for non-epistemic values has also expanded". She suggests, that is, that even if we manage to show that value judgments do not necessarily operate at the stage of hypothesis acceptance or rejection, we will still need to show that they do not necessarily operate at the three lower stages.

But the problem with Douglas's suggestion is that the inductive risk that is present at the three lower stages adds up to the total risk of the null hypothesis in accordance with the laws of probability (cf. Schurz, 2013, p. 330). This means that the total risk of the null hypothesis must not be lower than the inductive risk that is present at any of the three lower stages. In order to show that value judgments do not necessarily operate in the context of justification, we accordingly need to show that they do not necessarily operate at the stage of hypothesis acceptance or rejection; we do not need to show that they do not necessarily operate at any of the three lower stages.

Douglas (2007, pp. 9-10) believes that two problems confront the idea that the inductive risk that is present at the three lower stages adds up or "propagates" to the total risk of the null hypothesis. The first problem is that it can be difficult to estimate precisely the inductive risk that is present at any of the three lower stages; the second problem is the "more fundamental" problem that in order to estimate inductive risk, the scientist needs to make a probabilistic estimate, and that making such an estimate leads into a regress that is infinite unless at one point, the scientist accepts the hypothesis that the probability of another

hypothesis is *P*. Douglas (2007, p. 10) indicates that the second “more fundamental” problem “is first discussed in Rudner 1953”. It is arguably the problem that Rudner (1953, p. 4) describes when anticipating Jeffrey’s objection to premise (1) of AIR. I am going to deal with that second problem in the following section.

With respect to the first problem I would like to point out that in cases, where it is difficult to estimate precisely the inductive risk that is present at the second lower stage, scientists can work with upper or lower bounds of that risk. Consider the hypothesis that dioxins don’t cause liver cancer and the stage of expert pathologists judging whether a tissue sample has a cancerous lesion or not. In the presence of many borderline cases, it is difficult to estimate precisely the inductive risk of judging non-cancerous lesions to be malignancies, or of judging malignancies to be non-cancerous. But it is relatively easy to state upper and lower bounds of that risk. The upper (lower) bound of the inductive risk of judging non-cancerous lesions to be malignancies corresponds to the extreme case in which the pathologist judges all borderline cases to be malignancies (non-cancerous lesions). Similarly, the upper (lower) bound of the inductive risk of judging malignancies to be non-cancerous corresponds to the extreme case in which the pathologist judges all borderline cases to be non-cancerous lesions (malignancies).

In Douglas’s analysis, the pathologist targets the upper (lower) bound of the inductive risk of judging non-cancerous lesions to be malignancies and the lower (upper) bound of the inductive risk of judging malignancies to be non-cancerous if she prioritizes the interests of the general public (industry). In section 5, however, I am going to explain why the pathologist who prioritizes neither the interests of the industry nor those the general public should target the lower bound of the inductive risk of judging non-cancerous lesions to be malignancies and the upper bound of the inductive risk of judging malignancies to be non-cancerous. Douglas’s discussion of experiments that use laboratory animals to find out whether dioxins cause liver cancer nicely illustrates the various stages of hypothesis confirmation in which inductive risk is present. But it fails to demonstrate that we need to show that value judgments do not necessarily operate at the three lower stages if we manage to show that they do not necessarily operate at the upper stage.

3. Hypothesis acceptance or probability assignment?

Jeffrey’s objection to premise (1) of AIR says, to repeat, that “the activity proper to the scientist is the assignment of probabilities [...] to [...] hypotheses”, and not the acceptance or rejection of hypotheses. There are two criticisms that have been raised to this objection. Rudner (1953, p. 4) expresses the first criticism when anticipating Jeffrey’s objection. The criticism says that assigning probability P_0 to hypothesis H_0 is the same as accepting the

hypothesis H_1 that the probability of H_0 is P_0 , and that accepting H_1 presupposes value judgments. Jeffrey (1956, p. 246) responds that assigning probability P_0 to H_0 is *not* the same as accepting H_1 . But Douglas (2009, pp. 53-4) thinks that Jeffrey's response leads into a vicious regress of assigning probabilities to probability assignments.

It is by no means easy to state exactly the conditions under which a regress is vicious (cf. Priest, 2014, pp. 186-7). But in what follows, I shall assume that a regress is vicious if it is either infinite or circular. I will deal with infinity first and circularity second. The infinite regress of assigning probabilities to probability assignments looks as follows:

$$P(H_0) = P_0, P(P(H_0) = P_0) = P_1, P(P(P(H_0) = P_0) = P_1) = P_2 \dots \textit{ad infinitum}.$$

Instead of accepting the hypothesis that $P(H_0) = P_0$, we assign another probability P_1 to the hypothesis that $P(H_0) = P_0$, another probability P_2 to the hypothesis that $P(P(H_0) = P_0) = P_1$ and so on. If the regress is infinite and Jeffrey's response leads into this regress, then his response is arguably unjustified. But does it lead into this regress? I think that the answer is negative.

I think it is negative because a regress of assigning probabilities to probability assignments remains an empty formalism unless it is interpreted as an epistemic regress, and because in scientific practice, an epistemic regress is unlikely to be infinite. I want to demonstrate the unlikely infinity of the epistemic regress for the case of the type of experiments that Douglas discusses. The scientific practice of these experiments is that of statistical hypothesis testing, and in statistical hypothesis testing, the epistemic regress is likely to be finite. In order to see that, consider a variant of the hypothesis test that at least one of the studies that Douglas cites (Kociba et al, 1978) conducts: a variant of a test with two independent samples and dichotomous outcomes. The test proceeds in 3 or 4 steps, depending on whether you think that the genuine task of the scientist is to assign probabilities to hypotheses, or to accept or reject hypotheses.

In step 1, an experiment is designed in the way that Douglas describes. Let us assume that a random sample of $n = 200$ is drawn from a population of lab rats, that by some randomizing procedure, researchers assign $n_1 = 100$ rats to a treatment group and $n_1 = 100$ rats to a control group, and that the rats in the treatment group are exposed to dioxin, while the rats in the control groups remain non-dosed. Let us further assume that pathologists find malignancies in tissue samples of 25 of the rats in the treatment group and malignancies in tissue samples of only 15 of the rats in the control group, i.e. that the proportion of rats with malignancies in the treatment group is 0.25, while the proportion of rats with malignancies in the control group is 0.15. The following table summarizes the design of the experiment:

	n	N° of rats with	Proportion
--	-----	-----------------	------------

		malignancies	
Treatment group	$n_1 = 100$	$x_1 = 25$	$p_1 = 0.25$
Control group	$n_2 = 100$	$x_2 = 15$	$p_2 = 0.15$

In step 2 of the test, researchers assume that the distribution of p_1 and p_2 approximates standard normality. That assumption is likely to be satisfied because the individuals in the treatment and control groups form independent random samples (they have been assigned to a treatment and control group by some randomizing procedure), and because the two conditions for standard normality approximation, i.e. $p_1(1 - p_1)n_1 > 9$ and $p_2(1 - p_2)n_2 > 9$, are satisfied (we have $p_1(1 - p_1)n_1 = 16$ and $p_2(1 - p_2)n_2 = 12.75$).

In step 3 of the test, the Central Limit Theorem is used to calculate the likelihood of the null hypothesis, i.e. the probability of observing $p_1 - p_2 > 0$ if in fact $p_1 - p_2 = 0$, where $p_1 - p_2 = 0$ is selected as null hypothesis H_0 ('dioxins don't cause liver cancer') and $p_1 - p_2 > 0$ as alternative hypothesis. When using the Central Limit Theorem to calculate the likelihood of H_0 , we use the following test statistic:

$$z = (p_1 - p_2) \cdot [p(1-p)(1/n_1 + 1/n_2)]^{-1/2},$$

where $p = (x_1 + x_2) / (n_1 + n_2)$. The likelihood of H_0 is then given as $P(z > 1.754) = 1 - 0.96 = 0.04$, since in the standard normal distribution table, $z = 1.754$ corresponds to a probability of 0.96. Those who think with Jeffrey that the genuine task of the scientist is to assign probabilities to hypotheses, form the prior probability of H_0 and insert that probability together with the likelihood of H_0 into Bayes' theorem to calculate the posterior probability of H_0 . Those who think that the genuine task of the scientist is to accept or reject hypotheses proceed to step 4.

In step 4, H_0 is accepted or rejected in accordance with a decision rule that depends on the maximum probability α of committing a type I error of rejecting H_0 if it is true. If, for instance,

- $\alpha = 0.05$, the decision rule will require rejection of H_0 if $z > 1.645$ because in the standard normal distribution table, $z = 1.645$ corresponds to a probability of $1 - 0.05 = 0.95$.
- $\alpha = 0.01$, the decision rule will require rejection of H_0 if $z > 2.331$ because in the standard normal distribution table, $z = 2.331$ corresponds to a probability of $1 - 0.01 = 0.99$.

For the experiment designed in step 1, this means that H_0 will be rejected if $\alpha = 0.05$ and accepted (or not rejected) if $\alpha = 0.01$ (because $z = 1.754$). Following Douglas's analysis, one may say that researchers prioritizing the interests of the general public tend to select a high value for α , while researchers prioritizing the interests of the industry tend to select a low value. I'm going to argue in section 5, however, that researchers who accept or reject hypotheses independently of the value judgments that they happen to endorse, likewise tend to select a low value for α . And I will explain in the following section why researchers don't

need to determine a maximum probability β of committing a type I error of accepting H_0 if it is false.

There is arguably an epistemic regress departing from step 3 of the test, proceeding to step 2 and arriving at step 1. It is also possible to understand the regress as interpreting the regress of assigning probabilities to probability assignments: P_0 is the posterior probability of H_0 ; P_1 ('the probability that P_0 is the probability of H_0 ') is the prior probability of H_0 and the probability that p_1 and p_2 are approximately normally distributed; and P_2 ('the probability that P_1 is the probability that P_0 is the probability of H_0 ') is the probability that non-cancerous (cancerous) lesions have not been judged to be malignancies (non-cancerous), that potential confounders are distributed equally over the treatment and control group and so on. Estimations of these probabilities may even be quite precise: there are empirical procedures that can be used to test for the distribution of potential confounders over treatment and control groups; the probability that p_1 and p_2 are approximately normally distributed is quite high because the two conditions for normality approximation are satisfied; P_0 is provided with a specific value; and as noted in the previous section, there are upper and lower bounds to the probability that non-cancerous (cancerous) lesions have not been judged to be malignancies (non-cancerous).

But the important point is that the regress is not infinite: it ends with $P(P(P(H_0) = P_0) = P_1) = P_2$. There is no further probability P_3 that P_2 is the probability that ... There is also no reason why researchers should be required to accept the hypothesis that the probability that non-cancerous (cancerous) lesions have not been judged to be malignancies (non-cancerous), or that potential confounders are distributed equally over the treatment and control group. They assign a probability to that hypothesis (possibly in an informal way) and do not need to do anything else. Thus, Jeffrey's response does not lead into an infinite regress of assigning probabilities to probability assignments.

How about the circular variant of that regress? It looks as follows:

$$P(H_0) = P_0, P(P(H_0) = P_0) = P_1, P(P(P(H_0) = P_0) = P_1) = P_2, P(H_0) = P_0$$

Note that $P(H_0) = P_0$ shows up at the beginning and the end of the regress. Translated to hypothesis testing, the regress is one in which an experimental result (the result of a certain probability of the null hypothesis) is justified in terms of a the assumption of the approximate normal distribution of p_1 and p_2 , and in which the experimental design is justified in terms of the experimental result. We cannot rule out that such circular reasoning ever occurs in hypothesis testing, but it is definitely not a norm. Thus, Jeffrey's response does not lead into a circular regress of assigning probabilities to probability assignments either.

The second criticism of Jeffrey's objection says that scientists *should* accept or reject hypotheses because they are responsible for the actions that are taken on the basis of these hypotheses (Douglas, 2016, pp. 614-5, Wilholt, 2009, p. 94). In order to refute that criticism, Schurz (2013, pp. 325-328; 2014, p. 77) distinguishes categorical and hypothetical acceptances and rejections of hypotheses and argues that scientists (should) accept or reject hypotheses hypothetically, and not categorically, if they accept or reject hypotheses at all. He illustrates his distinction by means of a seismological example: if hypothesis H says that there won't be an earthquake of magnitude 6 or higher (in a specified area within the next few days), and if seismologists assign a probability of 0.95 to H , then they accept H

- categorically if they inform citizens that there won't be an earthquake of magnitude 6 or higher (if they sound the all-clear).
- hypothetically if they inform citizens about the cost-utility trade off they face (if they say quite literally that citizens should evacuate their homes if they find that the cost of remaining in their homes during an earthquake is at least 20 times as high as the cost of evacuating their homes temporarily when there is no earthquake).

Schurz (2013, p. 326) then mentions the legal trial after the earthquake in L'Aquila in 2009 as a case in point. The trial ended with a verdict that sentenced six seismologists for six years in prison for sounding the all-clear immediately before the earthquake. The verdict was later overturned: five of the seismologists were acquitted and only one of them, whose prison sentence was reduced, was found guilty of sounding the all-clear. But for Schurz, the trial is a case in point because it shows that scientists should *not* accept hypotheses categorically. The one seismologist, who was sentenced to prison, has indeed brought guilt upon him if it is true that he accepted the hypothesis that there won't be a major earthquake in L'Aquila in April 2009 categorically, i.e. if he sounded the all-clear (as it seems he did). He should have been acquitted together with his colleagues, by contrast, if he had accepted that hypothesis only hypothetically, i.e. if he had spoken of a low probability of that hypothesis, and if he had informed citizens of the cost-utility trade off that they faced.

I conclude that Jeffrey's objection remains unscathed: it is not obvious why Jeffrey's response to Rudner's criticism should lead into a vicious regress, and the claim that scientists should accept or reject hypotheses is too strong. This conclusion is not meant to imply that the genuine task of the scientist is to assign probabilities to hypotheses, and not to accept or reject these hypotheses. As I pointed out in the introduction, I prefer to remain agnostic about the two alternatives. But what the conclusion implies is that AIR is potentially unsound because premise (1) is possibly false.

4. Epistemic purity I: the Higgs boson discovery

Levi's objection says, to repeat, that there is an ambiguity in premises (2) and (3) of AIR: that the decision referred to in (2) is a decision about what to believe, while the decision referred to in (3) is a decision about how to act, that only the latter presupposes value judgments, and that the scientist qua scientist only needs to decide what to believe. Levi (1962, p. 48) argues that decisions about what to believe precede the acceptance or rejection of hypotheses in a non-behavioral sense, while decisions about how to act precede the acceptance or rejection of hypotheses in a behavioral sense. Thus, while Jeffrey, Douglas, Wilholt, and Schurz all understand (categorical) acceptances of hypotheses as acceptances in a behavioral sense, Levi (1960, p. 349) allows for the possibility of "accepting H in an 'open-ended' situation where there is no specific objective".³

Levi (1962, p. 49) understands a scientist who accepts or rejects hypotheses in a non-behavioral sense as one who seeks "the truth and nothing but the truth": as one who selects the true proposition from a set of competing possible propositions on the basis of the relevant evidence. Levi (1962, p. 51) points out that two constraints operate in the search for truth and nothing but the truth. The first ("hypothesis impartiality") requires that (a) the scientist do not prefer that any proposition from a set of competing propositions be true rather than another. The second ("error impartiality") requires that (b) she do not regard any possible error as being more serious than another.⁴ I am going to deal with error impartiality first, and with hypothesis impartiality second.

Remember that Douglas points out that significance level α for the maximum probability of committing a type I error and significance level β for the maximum probability of committing a type II error trade off each other, and that there seems to be no way to determine α and β without passing a value judgment about the respective consequences of selecting a high α or β . It seems impossible that a researcher can exactly balance these consequences and regard both errors as being equally serious. So how does Levi argue in favor of the possibility of error impartiality?

His argument involves the proposal that outcomes that fall outside the critical region should lead to suspension of judgment rather than acceptance of the null hypothesis, where the critical region is the region of test statistic values that result in rejection of the null hypothesis (in the example of section 3, this is the region where $z > 1.645$ if $\alpha = 0.05$ and $z > 2.331$ if $\alpha = 0.01$). Under this proposal, the result of rejecting a true null hypothesis is a mistake, while the result of suspending judgment about a false null hypothesis is not. Scientists may accordingly take type I errors more seriously than type II 'errors' without violating error

³ Levi (1960, p. 349) in fact reads Jeffrey as allowing for both possibilities (of accepting H in a behavioral and a non-behavioral sense) as well.

⁴ I borrow the terms "hypothesis impartiality" and "error impartiality" from Staley (2017, p. 12).

impartiality (cf. Levi, 1962, pp. 62-3). Levi points out that significance level α remains a matter of choice on the part of the investigator. But he also believes that α serves as “a rough index of the degree of caution exercised in a search for truth” (Levi, 1962, p. 63). He doesn’t think that this index presupposes value judgments.

One might respond that Levi’s proposal only pushes back the problem of selecting adequate significance levels to the stage of selecting the null hypothesis. In Douglas’s example, high α will lead to an over-regulation (under-regulation) of the dioxin-producing parts of the industry if the null hypothesis says that dioxins don’t cause (cause) liver cancer. If it is impossible for the scientist to remain impartial about null hypotheses and their alternatives, it will be impossible for her to remain impartial about errors. So how does Levi argue for the possibility of hypothesis impartiality?

His argument involves a well-known doctrine of classical hypothesis testing that Levi (1962, p. 62) quotes when saying that “the null hypothesis is to be selected in such a way that type I error will be more serious than type II error”. What needs to be understood, however, is that the seriousness in question is in the eye of the truth-seeking scientist. In the eye of that scientist, type I error (of rejecting a true null hypothesis) will be more serious than type II error (of accepting a false null hypothesis) as long as rejecting the null hypothesis amounts to a scientific discovery.

Classical hypothesis testing can be employed to test for all sorts of hypotheses: the hypothesis that a particular consignment of tulip bulbs contains 40 percent of the yellow- and 60 percent of the red-flowering sort, the hypothesis that the prevalence of cardiovascular disease among non-smokers in a given population is 10 percent and so on. In such cases I am unsure whether the truth-seeking scientist can select the null hypothesis in such a way that type I error will be more serious than type II error. But as long as rejecting a hypothesis amounts to a scientific discovery, the scientist needs to select that hypothesis as null hypothesis. The reason is that future research will build on that discovery: if the entity discovered is a causal relation, scientists will investigate the underlying mechanism; if the entity is a new particle, scientists will investigate its properties and so on. If rejecting the null hypothesis amounts to a scientific discovery, type I error will be more serious because future research would be misguided if a true null hypothesis were rejected.

Wilholt (2009, pp. 95-6) claims that Levi’s conception of a decision about what to believe “presupposes a sense of purity of epistemic activity that is exaggerated and unrealistic”. In the present and following section, I will concede that this claim is roughly correct. I will also argue, however, that what contaminates the epistemic activity of deciding what to believe is not necessarily a value judgment. In the rest of the present section, I will deal with the less

controversial case of the Higgs boson discovery. In the following section, I will turn to the more controversial case of science with clear non-epistemic impacts.

The Higgs boson discovery is the result of one comprehensive hypothesis test relying on the following statistical model (cf. van Dyk, 2014, p. 55):

$$N_{msc} \sim \text{Poisson}[\beta_{sc}(\theta_{sc}, m) + \kappa_{sc}(\phi_{sc}, m)\mu]$$

I will first explain the variables and parameters and then their subscripts. N is the number of observed events. An event is a proton-proton collision produced by the Large Hadron Collider (LHC) near Geneva. A proton-proton collision results in a trajectory of final-state particles. Particle detectors identify these particles by determining their momenta and/or energy. At the LHC, there are seven particle detectors. Two of them, ATLAS and CMS, were involved in the discovery of a Higgs boson. They identified this particle primarily through two decay channels: a Higgs boson decay into two photons ($H \rightarrow \gamma\gamma$) and a Higgs boson decay into two Z bosons ($H \rightarrow ZZ$), each of which decays, in turn, into two leptons (either electrons or muons).

The LHC produces millions of events per second but not of all of these are “observed”. The particle detectors have triggers that make very fast decisions as to whether an event is interesting or uninteresting, where an event is uninteresting if it involves well-understood physics. Rather than storing all events, the particle detectors save only those events that the triggers decide are interesting; these events amount to approximately 100 per second. Although this is a small fraction of the total events, it could still result in 10^{10} saved events for each experiment over the expected 15-year life span of the LHC. These saved events are the observed events. Their distribution is a Poisson distribution because counting the observed events is a Bernoulli trial (a random trial with exactly two possible outcomes: Higgs boson decay and no Higgs boson decay) with very large N and extremely low probabilities.

$\beta(\theta, m)$ models the “expected background count”: the number of events that can be expected to occur if all we have is well-understood physics, i.e. if the null hypothesis is true (the number of particle decays that do not represent Higgs boson decays).

$\kappa(\phi, m)$ models the “expected Higgs boson count”: the number of events that can be expected to occur if in addition to well-understood physics, there is new physics, i.e. if the null hypothesis is false (the number of particle decays that represent Higgs boson decays).

θ and ϕ are (vectors of) nuisance parameters (i.e. of parameters not primarily of interest: e.g. variances if the mean is a parameter of interest).

m is the Higgs mass associated with a specific “bin”. A bin is a potential Higgs mass on a fine grid of values of m_H (the unknown Higgs mass). Once a Higgs boson is discovered, its mass

can be estimated by including m_H in the model. Estimations indicate that the actual Higgs mass lies in the mass region around 126 GeV ($126 \cdot 10^9$ electronvolts).

μ is signal strength: the strength with which a particle decay “signals” its being a Higgs boson decay. Signal strength is a function of the (unknown) Higgs mass: the signal is strongest near the actual Higgs mass. Signal strength is defined so that $\mu = 0$ corresponds to the “background only hypothesis”, i.e. to the hypothesis that the observed particle decays do not represent Higgs boson decays, and $\mu = 1$ to the hypothesis that the observed particle decays represent Higgs boson and other particle decays. This allows $\mu = 0$ to serve as the null hypothesis subjected to significance testing.

I now turn to the subscripts of the variables and parameters. The subscripts indicate that the choice of both the background and Higgs boson models (their parameters and functional forms) is different for each category (stratum) s of (decay) channel c . The choice is different because once the particle detectors have saved the events that their triggers decide are interesting, events are “cut” within each decay channel, and because the events that survive the cuts are “stratified” into relatively homogenous categories: into categories with homogenous signal-to-background ratios and invariant mass resolutions.

The cuts aim to focus the analysis on a subset of events wherein new physical particles are more likely to be observed. The fraction of events that survive the cuts and involve new physics can be as low as 10^{-8} . In the actual Higgs boson discovery there were only a few hundred events in the two primary decay channels that could be associated with a Higgs boson decay. Stratification aims to increase the statistical power for identifying possible excess events above background that are due to new physics.

It is not entirely clear whether particle physicists can be said to accept the hypothesis that there are Higgs bosons. At least in official statements, they prefer to say that there is “conclusive evidence for the discovery of a new particle”, and that this evidence “is compatible with the hypothesis that the new particle is the Standard Model Higgs boson” (ATLAS collaborators, 2012, p. 15). But apart from these statements, which might simply express an abundance of caution, the Higgs boson discovery seems to exemplify a decision about what to believe rather well.

Particle physicists selected $\mu = 0$ as null hypothesis, because they believed that type I error (of rejecting a true hypothesis) will be more serious than type II error (of accepting a false hypothesis) if the null hypothesis says that $\mu = 0$. They believed that type I error will be more serious because future research would be misguided if $\mu = 0$ were rejected if true. At the same time, they selected a high “degree of caution” to minimize the risk of rejecting a true

null hypothesis: 6σ and 5σ in the ATLAS and CMS experiments, respectively, where a level of 5σ corresponds to a probability of about 1 in 3.4 million.

There have been several reasons for selecting a significance level of 5σ or higher. A purely conventional reason is that 5σ is the significance level that editors of particle physics journals generally require to claim a detection. Dawid (2015, pp. 79-80) cites a purely epistemic reason for selecting 5σ : the attempt to keep in check the so-called look-elsewhere effect, i.e. the problem that the probability of $\mu = 0$ that can be calculated for *any* category s of channel c (the “global” p -value) is greater than the minimum probability of $\mu = 0$ that can be calculated for a specific category s of channel c (the “local” p -value). Van Dyk (2014, p. 54) mentions a pragmatic reason for selecting 5σ : the attempt to account for model misspecifications that, given the large number of models for each category s of channel c , seem quite likely. Staley (2017b, p. 357) mentions another pragmatic reason: a “consideration of both the negative consequences of an erroneous discovery claim and the value for the further pursuit of inquiry of a correct discovery claim.”

It is true that only one of these reasons is purely epistemic. One may accordingly think that even in the case of the Higgs boson discovery, a decision about what to believe “presupposes a sense of purity of epistemic activity that is exaggerated and unrealistic”. But what contaminates epistemic activity in the case of the Higgs boson discovery is not a value judgment, but a bunch of conventional or pragmatic reasons. A good way to think about the difference is in terms of conditionals, the consequents of which make reference to methodological decisions (select $\mu = 0$ as null hypothesis, select 5σ or higher as significance level and so on). It is only in the case of value judgments that the antecedents refer to valuations of the utility of specific individuals or groups. In the case of conventional or pragmatic reasons, the antecedents make reference to technical goals (that of keeping the look-elsewhere effect in check, that of accounting for model misspecifications and so on).

One may of course object that technical goals are made explicit in official communiqués, while the utility valuations determine the methodological decisions as a matter of fact. One may argue with Staley (2017b, p. 369), for instance, that a false discovery claim would have been “tremendously embarrassing” for the physicists involved in the Higgs search, and that the 5σ significance level was selected to avoid that potential embarrassment.⁵ But the objection is not only “speculative” (as Staley himself admits); it also ignores that the wider community of scientists would not accept methodological decisions if they were taken to

⁵ Staley (2017b, p. 370) mentions further utility valuations, e.g. the anticipation of “negative consequences for the funding of HEP” and “a broader sense of responsibility toward the public perception of science in general”. Notice that he refers to both technical goals and utility valuations as “pragmatic”, while I reserve the term for technical goals.

increase the utility of specific individuals or groups. It ignores, for instance, that the majority of physicists not involved in the Higgs search would not accept the selection of 5σ as significance level if that selection would not allow for the achievement of specific technical goals.⁶

Staley (2017a, pp. 17-8; 2017b, p. 368) states that Levi's view about significance level α seems close to the view expressed by Douglas, and that his (Staley's) point that 5σ has been selected for a pragmatic reason resembles AIR. This statement seems to suggest that a decision about what to believe is impossible without value judgments even in the case of the Higgs boson discovery. It is important to understand, however, that this impossibility does not follow from Staley's analysis of the Higgs boson discovery. Staley (2017, p. 23) concedes that "the negative consequences of an erroneous discovery claim and the value for the further pursuit of inquiry of a correct discovery claim" can be regarded as epistemic values in a sense proposed by Steel. Staley prefers to think of these values as pragmatic, but what is clear is that they do not qualify as non-epistemic in the sense of value judgments. Thus even under Staley's analysis, the Higgs boson discovery exemplifies a decision about what to believe on the part of physicists: a decision that is epistemically pure in the sense of not depending on value judgments, but not epistemically pure in the sense of not depending on pragmatic considerations.

5. Epistemic purity II: science with non-epistemic impact

The more interesting question is, of course, whether science with clear non-epistemic impact (i.e. with impact on the utility of specific individuals or groups) may likewise exemplify a decision about what to believe that is epistemically pure in a similar sense. I think that the answer is positive, and I would like to return to Douglas's discussion of lab rat experiments to defend that answer. In Douglas's discussion of experiments that use lab rats to find out whether dioxins cause liver cancer, selecting high α (β) and low β (α) and judging most or all borderline cases to be malignancies (non-cancerous) leads to an excess of false positives (negatives) and an overregulation (under-regulation) of the dioxin-producing parts of the industry if the null hypothesis says that dioxins do not cause (cause) liver cancer. To "accept or reject" a hypothesis means to "accept or reject" the hypothesis in a behavioral sense: to act on the basis of that hypothesis, i.e. to (under- or over-) regulate the dioxin producing parts of the industry. Scientists will accordingly select high α (β) and low β (α) and judge most or all borderline cases to be malignancies (non-cancerous) if they prioritize the interests

⁶ What complicates matters a little is that the suitability of methodological decisions for achieving technical goals can be questioned (cf. van Dyk, 2014, p. 54). But the fact remains that the wider community of scientists would not accept a methodological decision unless it allows for the achievement of technical goals.

of the general public (the industry) and the null hypothesis says that dioxins *do not* cause liver cancer; they will select high β (α) and low α (β) and judge most or all borderline cases to be non-cancerous lesions (malignancies) if they prioritize the interests of the general public (the industry) and the null hypothesis says that dioxins *cause* liver cancer.

What if “accepting or rejecting a hypothesis” is understood in a non-behavioral sense? Then the acceptance or rejection of the hypothesis is preceded by a decision about what to believe that does not rely on judgments about the respective merits of under- or overregulating the dioxin-producing parts of the industry. Is it possible for a scientist to take such a decision in an experiment that uses lab rats to find out whether dioxins cause liver cancer? I don’t see why the answer should be negative. The scientist will conduct a hypothesis test that is similar to the one presented in section 3.

She will first design an experiment like the one that Douglas describes. She will, secondly, make assumptions about the underlying probability distribution. Third, she will select the hypothesis that dioxins don’t cause liver cancer as null hypothesis because she will judge that committing a type I error of rejecting a true null will be more serious than committing a type II error of accepting a false null if the null hypothesis says that dioxins don’t cause liver cancer. She will judge that committing a type I error will be more serious because as a scientist seeking “the truth and nothing but the truth”, she will understand that rejecting the hypothesis that dioxins don’t cause liver cancer amounts to the discovery of a causal relationship, that future research should investigate the mechanism underlying that relationship, and that future research would be misguided if the hypothesis were rejected if it were true.

The scientist will finally judge most or all borderline cases to be non-cancerous lesions and select low α because as a scientist seeking “the truth and nothing but the truth”, she aims to minimize the risk of rejecting the null hypothesis if it is true. She will not explicitly select any value for β because as a scientist seeking “the truth and nothing but the truth”, she understands that test statistic values that fall outside the critical region should lead to suspension of judgment rather than acceptance of the null hypothesis. She will reject the null hypothesis (and decide to believe that dioxins cause liver cancer) if her test statistic value falls inside the critical region, but she will suspend judgment otherwise.

Her judgment of borderline cases and her selections of the null hypothesis and significance level α depend on a pragmatic “consideration of both the negative consequences of an erroneous discovery claim and the value for the further pursuit of inquiry of a correct discovery claim.” It is important to understand, however, that her judgment of borderline cases and her selections of the null hypothesis and significance level α remain unaffected by

judgments about the respective merits of under- or overregulating the dioxin-producing parts of the industry. She takes a decision about what to believe independently of any of the value judgments that she happens to endorse.

There are two objections that one might feel should be raised to the idea that in experiments using lab rats to find out whether dioxins cause liver cancer, scientists can take decisions about what to believe independently of value judgments. The first objection seems to confirm the Marxist suspicion that scientists are “sycophants of capital”. It says that the scientists who (only seem to) take decisions about what to believe independently of value judgments in fact serve the interests of the industry involuntarily and unconsciously. Since α (the maximum probability of committing a type I error) and β (the maximum probability of committing a type II error) trade off each other, low α will lead to high β and, consequently, an excess of false negatives and an under-regulation of the dioxin-producing parts of the industry if the null hypothesis says that dioxins don’t cause liver cancer.

But the objection loses sight of the fact that one would likewise have to say that in science with clear non-epistemic impact (such as pharmacology), scientists who (only seem to) take decisions about what to believe independently of value judgments sometimes serve the interests of the general public involuntarily and unconsciously. Consider, for instance, the hypothesis that selective serotonin reuptake inhibitors (SSRIs) have a favorable risk-benefit profile in anti-depression treatment.⁷ The scientist seeking “the truth and nothing but the truth” will select the hypothesis that SSRIs do *not* have a favorable risk-benefit profile as null hypothesis because she understands that rejecting that hypothesis amounts to a scientific discovery, and that committing a type I error will be more serious than committing a type II error if the null hypothesis says that SSRIs do *not* have a favorable risk-benefit profile. She will then select low α because she aims to minimize the risk of committing a type I error. Since α and β (the maximum probability of committing a type II error) trade off each other, low α will lead to high β and, consequently, an excess of false negatives and a restriction of trade in SSRIs, which benefits the general public if the risk-benefit profile of SSRIs turns out to be unfavorable (as it in fact did).

But the objection also loses sight of the fact that under Levi’s proposal, α and β do not trade off each other (because a type I error is a mistake, while a type II ‘error’ is not), that the scientist seeking “the truth and nothing but the truth” will accept the alternative hypothesis only in a non-behavioral sense if she manages to reject the null hypothesis, and that she will never accept the null hypothesis (she will only suspend judgment about whether or not dioxins cause liver cancer if she doesn’t manage to reject the null hypothesis). It is

⁷ I borrow the example from Brown (2008, p. 193).

accordingly false to say that low α will lead to high β and an under-regulation of the dioxin-producing parts of the industry, or more generally, that significance level selections will lead to any type of policy measure.

Policy measures will, in any case, remain in the hands of the authorities. If the scientist seeking “the truth and nothing but the truth” manages to reject the null hypothesis, she will accept the alternative hypothesis, but “accepting the alternative hypothesis” is not synonymous with “acting on the basis of it”. Acting on the basis of it will be left to the authorities who will (hopefully) regulate the dioxin-producing parts of the industry. If the scientist seeking “the truth and nothing but the truth” does not manage to reject the null hypothesis, she will not accept the null hypothesis, but suspend judgment about whether or not dioxins cause liver cancer. Action will again be left to the authorities who will under-regulate the dioxin-producing parts of the industry if they prioritize the interests of the industry and over-regulate the industry if they prioritize the interest of the general public. They might also be able, however, to balance the interests of the general public and the industry to a reasonable degree.

The second objection that one might feel should be raised to the idea that in experiments using lab rats to find out whether dioxins cause liver cancer, scientists can take decisions about what to believe independently of value judgments, is that such scientists are fictions. No scientist conducting experiments of this kind ever judges borderline cases or selects null hypotheses and significance levels independently of the judgments about the respective merits of over- and under-regulating the dioxin-producing parts of the industry. The objection is not only as speculative as the objection mentioned toward the end of the preceding section; it also misses its target. The idea is that in experiments using lab rats to find out whether dioxins cause liver cancer, scientists *can* take decisions about what to believe independently of value judgments. Even if they often fail to take such decisions, they are at least capable of taking them.

Let me try to explain why I think that it is important to insist that it is at least possible for them to take such decisions. Authors like Brown and Wilholt like to cite meta-studies that find that “only 5 percent of published reports on new drugs that were sponsored by the developing company gave unfavorable assessments”, while “38 percent of published reports were not favorable when the investigation of the same drugs was sponsored by an independent source” (Brown, 2008, p. 191); that “90% of government-funded experiments on low-dose exposure to bisphenol A reported significant [detrimental health] effects, while not a single industry-funded experimental study did so” (Wilholt, 2009, p. 93); and so on. What these meta-studies show is that in pharmacology, material or group interests (usually, but not necessarily, those of the pharmaceutical industry) often bias decisions about what to believe.

But scientists need to hold on to the value-free ideal if they wish to criticize what they perceive to be biased decisions, and if they wish to avoid being drawn into “critical interactions among scientists of different points of view” (Longino, 1996, p. 40). According to Longino (1996, p. 40), these interactions aim “to mitigate the influence of subjective preferences on [...] theory choice” and “to transform the subjective into the objective”. But perhaps one might also understand them as aiming to establish one set of value judgments as superior to another. Scientists who wish to criticize what they perceive to be biased decisions, and who wish to avoid “critical interactions”, will need to assume that it is possible to decide what to believe independently of material or group interests (or value judgments, more generally).

Philosophers who accept the conclusion of AIR feel compelled to believe that for science “that has clear non-epistemic impacts, being ‘value-free’ is not a laudable goal” (Douglas, 2000, p. 560), or that science is determined by conventional standards solving coordination problems (Wilholt, 2009, pp. 97-99). It is important to understand, however, that scientists won’t be able to criticize a pharmaceutical study for being biased by the material interests of the industry if being ‘value-free’ is not a laudable goal, and that what they criticize when criticizing a pharmaceutical study for being biased by material interests cannot be the violation of a conventional standard that is supposed to solve coordination problems. Material or group interests (or value judgments of various sorts) might, after all, bias the defense or implementation of that standard itself. Scientists need to hold on to the value-free idea if they wish to criticize a pharmaceutical study for being biased by the material interests of the industry, and if they want to avoid being drawn into critical interactions. In order to hold onto the ideal, they do not need to attain it as a matter of fact. But they need to be capable of attaining it, and AIR doesn’t manage to show that they are not capable of attaining it.

6. Conclusion

The argument from inductive risk, as developed by Rudner and others, famously concludes that the scientist qua scientist makes value judgments. In the preceding sections, I tried to show that trust in the soundness of the argument is overrated because philosophers who endorse its conclusion (especially Douglas and Wilholt) fail to refute two of the most important objections that have been raised to its soundness: Jeffrey’s objection that the genuine task of the scientist is to assign probabilities to (and not to accept or reject) hypotheses, and Levi’s objection that the argument is ambiguous about decisions about how to act and decisions about what to believe, that only the former presuppose value judgments, and that qua scientist, the scientist only needs to decide what to believe.

I pointed out that Douglas argues against Jeffrey's objection that it leads into a vicious regress of assigning probabilities to probability assignments, that Douglas and Wilholt argue against Jeffrey's objection that scientists *should* accept or reject hypotheses because they are responsible for the actions that are taken on the basis of these hypotheses, and that Wilholt (2009, pp. 95-6) argues against Levi's objection that his conception of a decision about what to believe "presupposes a sense of purity of epistemic activity that is exaggerated and unrealistic".

I then tried to show in section 3 that Jeffrey's objection is unlikely to lead into a vicious regress, and that (as Schurz suggests) scientists should accept or reject hypotheses only hypothetically (if they should accept or reject hypotheses at all). I tried to show in sections 4 and 5 that decisions about what to believe might not be epistemically pure in the sense envisaged by Levi, but that what supplements decisions about what to believe may be pragmatic considerations that do not necessarily involve value judgments.

Let me conclude by making a couple of remarks on what I did not try to show, and what I think would be worth investigating in the future. As I mentioned in the introduction, I did not try to defend either Jeffrey's or Levi's objection. I pointed out that a defense of both objections would be self-defeating because the view that the activity proper to the scientist is not the acceptance or rejection of hypotheses is incompatible with the view that the scientist qua scientist needs to decide what to believe. Thus an interesting question concerns the genuine task of the scientist: is it the assignment of probabilities, a decision about what to believe, or something else? The answer to this question depends on the respective positions of realism or antirealism that one is prepared to adopt (cf. footnote 2 above). So another interesting questions relates to the exact nature of these positions and to the reasons that speak against and in favor of these positions.

I did not try to show either that the scientist qua scientist does not make value judgments, or that value judgments do not necessarily operate in the context of justification. I tried to show that trust in the soundness of AIR is overrated, but I am fully aware that there are other arguments that aim to derive the same conclusion as AIR. Let me name one of these arguments: an argument that Longino has developed in a series of books and papers (especially in Longino, 1990), i.e. the argument from empirical underdetermination (AEU). AEU derives the conclusion that the scientist makes value judgments qua scientist from (a variant of) Quine's thesis of the underdetermination of scientific hypotheses by empirical evidence, i.e. from the thesis that scientific hypotheses are underdetermined by empirical evidence persistently and non-sporadically (cf. Norton, 2008, p. 20).

I think that Norton (2008, pp. 18, 19) is right when saying that “the underdetermination thesis is little more than speculation”, and that the question of the empirical underdetermination of scientific hypotheses should be decided “on a case-by-case basis”. But once we observe that a scientist accepts or rejects a scientific hypothesis, even though that hypothesis is empirically underdetermined (once we observe, for instance, that a scientist accepts the null hypothesis, even though outcomes fall outside the critical region), we will have reason to believe that the scientist makes a value judgment. Without further ado, we won’t be able to decide what value judgment the scientist makes, or whether the scientist makes the value judgment qua scientist. But we will observe the operation of a value judgment at the core of scientific activity, and an interesting question concerns possible methods of identifying that judgment.

Another interesting question relates to the ideal organization of scientific discourse in cases, in which the empirical under-determination of a scientific hypothesis is persistent and non-sporadic. As has been pointed out toward the end of the preceding section, Longino (1996, p. 40) recommends “critical interactions among scientists of different points of view”. She also proposes specific rules that should govern these interactions (cf. e.g. Longino 1996, p. 40). One may doubt, however, whether the rules are adequate, or whether the interactions really have the ability that Longino claims they have (“to mitigate the influence of subjective preferences on [...] theory choice” and “to transform the subjective into the objective”). One may, alternatively, recommend discussions among scientists (perhaps involving the general public) about the respective merits of various value judgments and propose rules that should govern these discussions. One may also question, however, whether these discussions will take us anywhere, and pray for empirical evidence to be forthcoming and to remove the underdetermination of the hypothesis in question.

References:

- ATLAS Collaboration (2012). Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. *Physics Letters B* 716, 1-29.
- Brown, J. R. (2008). The Community of Science®. In M. Carrier, D. Howard, & J. Kourany (Eds.), *The Challenge of the Social and Pressure of Practice: Science and Values Revisited*. Pittsburgh: University of Pittsburgh Press (189-216).
- Carnap, R. (1950). *Logical Foundations of Probability*. Chicago: University of Chicago Press.
- Churchman, C. W. (1948). Statistics, Pragmatics, Induction. *Philosophy of Science* 15, 249-268.
- Dawid, R. (2015). Higgs discovery and the look elsewhere effect. *Philosophy of Science* 82, 76-96.

- Douglas, H. (2000). Inductive Risk and Values in Science. *Philosophy of Science* 67, 559-579.
- Douglas, H. (2007). Rejecting the ideal of value-free science. In H. Kincaid, J. Dupré, & A. Wylie (Eds.), *Value-free Science? Ideals and Illusions*. Oxford: OUP (120-141).
- Douglas, H. (2009). *Science, Policy, and the Value-Free Ideal*. Pittsburgh: Pittsburgh University Press.
- Douglas, H. (2016). Values in Science. In P. Humphreys (Ed.), *The Oxford Handbook of Philosophy of Science*. Oxford: OUP (609-630).
- Hempel, C. G. (1949). Review of Churchman's *Theory of Experimental Inference*. *Journal of Philosophy* 46, 557-561.
- Hempel, C. G. (1965). Science and Human Values. In *Aspects of Scientific Explanation and Other Essay in the Philosophy of Science*. New York: The Free Press (81-96).
- Jeffrey, R. (1956). Valuation and Acceptance of Scientific Hypotheses. *Philosophy of Science* 23, 237-246.
- Kitcher, P. (2011). *Science in a Democratic Society*. Amherst: Prometheus.
- Kociba, R. et al. (1978). Results of a Two-Year Chronic Toxicity and Oncogenicity Study of 2, 3, 7, 8-Tetrachlorodibenzo-p-Dioxin in Rats. *Toxicology and Applied Pharmacology* 46, 279-303.
- Levi, I. (1960). Must the Scientist Make Value Judgments? *Journal of Philosophy* 57, 345-357.
- Levi, I. (1962). On the Seriousness of Mistakes. *Philosophy of Science* 29, 47-65.
- Levi, I. (1984). Messianic vs Myopic Realism. *Proceedings of the Biennial Meeting of the Philosophy of Science Association 1984*, 617-636.
- Longino, H. E. (1990). *Science as Social Knowledge*. Princeton, NJ: Princeton University Press.
- Longino, H. E. (1996). Cognitive and Non-cognitive Values in Science. In L. H. Nelson and J. Nelson (eds.), *Feminism, Science, and the Philosophy of Science*. London: Kluwer, pp. 39-58.
- Norton, J. D. (2008). Must Evidence Underdetermine Theory? In M. Carrier, D. Howard and J. Kourany (eds.), *The Challenge of the Social and Pressure of Practice: Science and Values Revisited*. Pittsburgh: University of Pittsburgh Press, pp. 17-44.
- Priest, G. (2014). *One*. Oxford: Oxford University Press.
- Rudner, R. (1953). The Scientists qua Scientists Makes Value Judgments. *Philosophy of Science* 20, 1-6.
- Schurz, G. (2013). Wertneutralität und hypothetische Werturteile in den Wissenschaften. In G. Schurz & M. Carrier (Eds.), *Werte in den Wissenschaften*. Frankfurt am Main: Suhrkamp (305-334).

- Schurz, G. (2014). *Philosophy of Science. A Unified Approach*. London: Routledge.
- Staley, K. W. (2017a). Decisions, Decisions: Inductive Risk and the Higgs Boson. In K. C. Elliott & T. Richards (Eds.), *Exploring Inductive Risk: Case Studies of Values in Science*. Oxford: OUP.
- Staley, K. W. (2017b). Pragmatic Warrant for Frequentist Statistical Practice: The Case of High Energy Physics. *Synthese* 194, 355-376.
- Van Dyk, D. A. (2014). The role of statistics in the discovery of the Higgs boson. *Annual Review of Statistics and its Applications* 2014(1), 41-59.
- Wilholt, T. (2009). Bias and Values in Scientific Research. *Studies in History and Philosophy of Science* 40, 92-101.