Is the value-free ideal of science untenable? Part I: Inductive risk

Duygu Uygun Tunç
University of Chicago

Mehmet Necip Tunç
Tilburg University

Correspondence regarding this manuscript should be addressed to uyguntunc@uchicago.edu

Draft Date: 04-21-2025
Word count: 11,335

Abstract

The inductive risk argument challenges the value-free ideal of science by asserting that scientists should manage the inductive risks involved in scientific inference through social values, which consists in weighing the social implications of errors when setting evidential thresholds. Most of the previous analyses of the argument fall short of engaging directly with its core assumptions, and thereby offer limited criticisms. This paper critically examines the two key premises of the inductive risk argument: the thesis of *epistemic insufficiency*, which asserts that the internal standards of science do not suffice to determine evidential thresholds in a non-arbitrary fashion, and the thesis of *legitimate value-encroachment*, which asserts that non-epistemic value judgments can justifiably influence these thresholds. A critical examination of the first thesis shows that the inductive risk argument does not pose a unique epistemic challenge beyond what is already implied by fallibilism about scientific knowledge, and fails because the mere assumption of fallibilism does not imply the untenability of value-freedom. This is demonstrated by showing that the way in which evidential thresholds are set in science is not arbitrary in any sense that would lend support to the inductive risk argument. Relatedly, value-laden strategies would likely jeopardize the emergence of a rational consensus by prematurely resolving scientific debates. A critical examination of the thesis of legitimate value-encroachment shows that incorporating social values into scientific inference as an inductive risk-management strategy faces a meta-criterion problem, and consequently leads to several serious issues such as wishful thinking, category mistakes in decision making, or Mannheim-style paradoxes of justification. Consequently, value-laden strategies for inductive risk management would likely weaken the justification of scientific conclusions.

Keywords: Values in Science, Value-free Ideal, Value Neutrality, Inductive Risk, Evidential Threshold, Pragmatic Encroachment

1. Introduction

Scientific justification is defeasible and even our best claims often carry some level of uncertainty. Since this is the case, scientists are faced with a rather fundamental problem when they are making inferences based on their empirical findings: How far should the uncertainty be reduced before accepting or rejecting scientific hypotheses? Additional evidence reduces uncertainty, but it cannot entail the sufficiency of this reduction. Therefore, the scientific decision to accept or reject a hypothesis inevitably involves extra-evidential standards. A centerpiece of scientific normativity is that scientific inference ought to be value-free, and therefore the only kind of value judgments that can be legitimately used to manage the risks pertaining to scientific error are those that are internal to science (epistemic/cognitive values or standards). This idea has been challenged by a variety of arguments, among which the inductive risk argument is arguably one of the strongest to date.

The inductive risk argument against the value-free ideal of science states that because scientific inferences are characteristically error-prone, the practical costs of making an error should influence the evidential standards for accepting or rejecting scientific claims (Douglas, 2000, 2009; Rudner, 1953). Douglas (2009, p. 97) expresses it in terms of individual scientific judgment: "scientists should [weigh the importance of] the potential social and ethical consequences of error . . . and set burdens of proof accordingly." [1] Rudner (1953, p. 2)'s original formulation reads:

> ...since no scientific hypothesis is ever completely verified, in accepting a hypothesis the scientist must make the decision that the evidence is sufficiently strong or that the probability is sufficiently high to warrant the acceptance of the hypothesis. Obviously our decision regarding the evidence and respecting how strong is "strong enough" is going to be a function of the importance, in the typically ethical sense, of making a mistake in accepting or rejecting the hypothesis.

This argument is commonly interpreted as the claim that scientists, as a general feature of scientific judgment, "have to make value-laden decisions about how much evidence to demand before they draw conclusions" (Elliott, 2022). Douglas (2009, p. 87) says explicitly that social values are required at the core of science, "not just as a matter of an accurate description of scientific practice, but as part of an ideal for scientific reasoning."

Several authors in various ways have forcefully criticized the inductive risk argument, such as rejecting the premise that scientists accept or reject hypotheses (Jeffrey, 1956), rejecting the assumption that if they do accept hypotheses they must do so without qualification, such as conditionalizing their claims or articulating uncertainties (Betz, 2013; Henschen, 2021), distinguishing between accepting a hypothesis as true and acting on its basis (Levi, 1960, 1962), further disambiguating the notion of 'acceptance' into distinct cognitive attitudes, such as endorsing, adopting, or holding a hypothesis (Lacey, 2015), arguing that value-free ideal is indispensable for the political legitimacy of science (Betz, 2013; Lusk, 2021), or that it is pursuitworthy because of its desirable epistemic consequences (Menon & Stegenga, 2023; Stegenga & Menon, 2023). Also, due to the emphasis on practical consequences, the argument from inductive risk can be said to pose a limited challenge to the value-free ideal of science, one that primarily concerns cases where (i) there are foreseeable practical consequences to accepting or rejecting scientific claims with a view to real world application (Elliott, 2011; McMullin, 1982), and (ii) withholding judgment (Giere, 2003) or deferral (Betz, 2013; Havstad & Brown, 2017) are not desirable. This reading of the inductive risk argument does not pose a substantial problem for the value-free ideal beyond the contexts of scientific policy advice (Steele, 2012) and possibly fast science (Stegenga, 2024). But, this limited formulation of the inductive risk overlooks the stronger argument that the error-prone nature of scientific conclusions requires some form of risk management, and that these risks cannot be managed solely in terms of epistemic values.

While we believe that most of these objections still

[1] To use Douglas' own example for illustration, toxicologists must decide what level of statistical significance to demand in order to conclude whether the chemical dioxin under investigation indeed increased cancer rates in animal experiments. Depending on where they set the significance level, they trade off the probability for false positive results with the probability for false negatives (commonly referred to as error rates), where a false positive means that experiments detect an increase in cancer rates where there is none (or detect a larger increase than the actual value), a false negative means a failure to detect a real increase in cancer rates (or detect a smaller increase than the actual value). Concluding the carcinogenic effects of dioxin in humans (by extrapolation from animal studies) has the foreseeable practical consequence of increased regulation of the chemical and thereby excess costs to the industry, where concluding the opposite will likely lead to weaker regulation and thereby costs related to public health. Douglas concludes that the toxicologists should decide the right balance between the error rates based on how they value these effects, as a function of their social and moral values as ordinary human beings or citizens.

stand, there has been almost no attempt [2] to take on and criticize the inductive risk argument as a generalized claim against the epistemic self-sufficiency of science–the capacity of science to manage inductive risks without appeal to anything other than its internal standards. [3]. While steelmanning the inductive risk argument, only a successful criticism of this central and generalized challenge can put this debate finally to rest. Arguably this lack is one of the reasons why the inductive risk argument against the value-free ideal still has an intuitive appeal for many. This paper attempts to take on this challenge by arguing that the inductive risk argument does not suggest the untenability of the value-free ideal of science any more than what can be inferred from the thesis of fallibilism about scientific knowledge. We cannot infer the untenability of the value-free ideal of science merely from fallibilism about knowledge, hence the inductive risk argument fails. Moreover, appeal to social values in judgments of evidential sufficiency would further undermine the justification of scientific inference, and thus make the problem (if any) simply worse.

As a generalized challenge to the value-neutrality of scientific judgment, the argument from inductive risk can be broadly characterized as an application of the *pragmatic encroachment thesis* in epistemology to scientific judgments of *evidential sufficiency*. [4] The argument can be reconstructed as having two key premises. The first one advances an analogue of the *threshold problem* for *fallibilism* about knowledge against the concept of evidential sufficiency in science. Fallibilism about knowledge says that knowledge can be achieved with less than fully conclusive justification. The threshold problem is about how to determine the level of justification that separates knowledge from lack thereof in a non-arbitrary way. Douglas (2009) expresses an analogous challenge for scientific knowledge by saying that there is no non-arbitrary, non-pragmatic way to decide when the evidence is sufficient to accept a scientific claim without considering social values. This gives us what we will call the *thesis of epistemic insufficiency*.

> **Epistemic insufficiency:** There is no epistemic basis on which a threshold of justification that will cover all instances of scientific knowledge can be determined.

This leads to the idea of science's self-insufficiency; its insufficiency to justify its core practice, scientific inference, by its internal standards alone. Douglas clearly puts forward this thesis, when she says that purely "internal standards" –i.e., standards that are "free of social and ethical values," such as methodological standards, theoretical virtues, or cognitive values— do not help to decide "what counts as sufficient evidence" in a non-arbitrary way (Douglas, 2017, also 2000; 2009). Hence, the internal standards are insufficient for scientific judgment regarding which propositions to accept.

The second key premise of the argument offers the consideration of social values as a solution to the problem of epistemic insufficiency, which is analogous to the *impurist* solution to the threshold problem: Whether a subject S knows a proposition p depends not only on epistemic factors such as evidence or reliability, but also on the stakes involved in S's practical reasoning situation, or how important the truth of p is to S (Fantl & McGrath, 2002, 2009; Stanley, 2005). Thus, the pragmatic encroaches on the epistemic. The pragmatic encroachment thesis states that practical factors, such as a subject's practical interests regarding a certain content or the stakes involved in falsely affirming or disaffirming that content, are relevant in determining the epistemic standards that must be met in order for a subject's belief to be sufficiently justified to constitute knowledge (see Kim, 2017). Applied to the scientific context, this can be seen as analogous to saying that social values are relevant to determining whether there is sufficient evidence to accept or reject a scientific hypothesis, which is equivalent to the inductive risk argument (see also Miller, 2024). This part of the argument gives us what

[2] With the possible exception of some discussions in Levi (Levi, 1960, 1962).

[3] Inductive risks, broadly construed, are the risks pertaining to (i) accepting a scientific claim as true when it is in fact false, and (ii) rejecting a scientific claim as false when it is in fact true (see also, Hempel, 1960)

[4] To revisit Douglas' dioxin research example, shifting the statistical significance level in either way directly affects how the theory choice between two alternative dose-response models is resolved, namely the threshold model and the linear extrapolation model. The threshold model states that the carcinogenic effect of dioxin starts at a certain dose of the chemical (the threshold), below which there is no such effect. The linear extrapolation model, on the other hand, states that the dose has a linear relationship with the carcinogenic effect, meaning that there is no dose threshold below which the chemical is completely non-carcinogenic. Where the statistical significance level is set affects both which data patterns will be considered a response and the shape of the dose-response curve (Douglas, 2000). What creates the theory choice situation in such a case is that the studies are not sufficiently powered (i.e., have sufficient sample size) to be able to generate the evidence that will distinguish which model or hypothesis is the correct one. Collecting larger samples has considerably higher economic costs, which might lead the researchers (or other parties such as funders, policy makers) to resolve (rather than 'solve') the theory choice situation by making value-laden judgments.

we will call the thesis of *legitimate value-encroachment*.

> **Legitimate value-encroachment:** Social values are legitimate determinants of where the threshold of justification for scientific knowledge is to be set.

In conjunction, these two premises are used to conclude that the value-free ideal of science is untenable, and only a value-laden science can legitimately manage inductive risks. This paper criticizes this conclusion by disentangling the inductive risk argument against the value-free ideal of science from the thesis of fallibilism about scientific knowledge. We argue that there is no additional epistemic challenge contained in the inductive risk argument which is not found in the thesis of fallibilism, and since the latter does not imply the untenability of the value-free ideal of science, neither does the inductive risk argument. Moreover, the positive part of the inductive risk argument which argues for a value-laden ideal for science has more serious problems of its own. Firstly, we reject that thresholds of scientific justification are arbitrary in several commonly understood senses of arbitrariness that would justify the inductive risk argument, and argue that purely epistemic considerations suffice to eliminate these. However, non-arbitrariness in these senses does not imply the *impossibility of rational disagreement in science*, that is disagreement between scientists on the basis of strictly epistemic considerations. This stronger thesis would require in fact to deny the fallibilist nature of scientific knowledge. Since we by default accept the fallibility of scientific knowledge, we also by default accept the possibility of rational disagreement in science. But neither the fallibility of scientific knowledge nor the consequent possibility of rational disagreement in science by themselves imply that scientific inference necessitates non-epistemic value judgments, or that it can be improved by these. If the critics of the value-free ideal hang their claim on the bare assumption of fallibilism, their reconstruction of the value-free ideal is nothing but a *reductio ad absurdum*, because it makes the tenability of the value-free deal dependent on an assumption of infallibilism.

If this is not true, the proponents of the inductive risk argument should explain what social value-judgments bring into the picture: If an epistemic problem is unsolvable, arguably it is not made *any less unsolvable* by introducing non-epistemic elements. There are strong reasons to think that any rational disagreement in science is transient and thus will eventually give way to rational consensus, because epistemic value-judgments that underlie disagreements are subordinated to higher-order shared epistemic values such as truth, and thus themselves are subject to epistemic evaluation. But even if this is but an unfounded hope, we are not justified to think that value-laden science would be any better, and there are good reasons to think that it would be much worse. This is mainly due to the fact that the justification of social values, unlike epistemic values, is itself subjective and thereby disagreements about social values prevent rational consensus.

The structure of the paper is as follows. In §2 we explain the context of use for evidential thresholds in science and the kind of scientific judgments they pertain to. In §3 we analyze several commonly understood senses of arbitrariness and show that scientific evidential thresholds are not arbitrary in any of these senses. We conclude that the inductive risk argument misconstrues the justification of evidential sufficiency criteria, particularly the statistical significance thresholds. In §4 we examine another possible interpretation of the thesis of epistemic insufficiency as the possibility of rational disagreement about the interpretation of evidence, and discuss how rational disagreements are resolved in science. Section §5 turns to the thesis of legitimate value-encroachment and tackles the epistemic problems that emerge if social values are used to set evidential thresholds. We argue that Douglas' distinction between legitimate and illegitimate uses of values in science fails to hold due to a paradox of meta-criterion. Section §6 distinguishes between judgments of evidential sufficiency in the scientific vs practical decision-making contexts and argues that there is no encroachment between the two contexts in either way. In §7 we argue that the thesis of legitimate value-encroachment suffers a meta-criterion problem of its own, and conclude that value-laden evidential sufficiency judgments cannot increase but only undermine the legitimacy of evidential sufficiency judgments.

2. Evidential thresholds and acceptance of scientific hypotheses

2.1. What is an evidential threshold in scientific inference?

Scientific (or theoretical) claims are on the level of phenomena, while empirical observations occur on the level of data. Phenomena are relatively stable features of the universe that are not dependent on particular observations, while the specific data that are observed depend on details of the experimental procedure and the measurement device (Woodward, 1989; Woodward, 2000). Consequently, although phenomena are detected by data, the data never directly or necessarily entail claims about phenomena (Bogen & Woodward, 1988), because there is always the possibility of error,

for example due to measurement or sampling (Woodward, 1989). Statistical error control is an integral part of scientific observation where data is known to contain non-negligible error(see also 'corrigible data' in Suppes, 1974). It is a theory of error which enables us to compare alternative scientific hypotheses in a way that the difference of evidential support cannot be attributed to variations due to random error (Mayo, 1996). To do this, scientists estimate a confidence interval and define an evidential threshold.[5] An evidential threshold can thus be seen as one kind of evidential sufficiency criterion. Rudner (1953) and Douglas (2000, 2009) address the most widely used criterion of evidential sufficiency in statistics, namely the type 1 error rates, or α-levels, for determining statistical significance.

2.2. Evidential sufficiency to answer which question? Acceptance of evidence vs acceptance of a theoretical claim

A statistical significance threshold is used for the acceptance of evidence, rather than the acceptance of a scientific claim into the canon. [6] Scientists use type 1 error rates (i.e., p-values) to decide whether the data expresses a singular fact (such as whether the patient outcomes in the experimental condition veritably differ from those in the control condition), not to decide between two substantive hypotheses (such as whether smoking causes cancer or not). This is the difference between a statistical hypothesis and a scientific claim. By testing a statistical hypothesis, scientists establish an evidential link between the data and the substantive hypothesis. The substantive hypothesis or the scientific claim is almost never justified on the basis of a statistical hypothesis-test alone, because finding a true positive signal in the data does not rule out all systematic sources of error, such as measurement invalidity, or sufficiently probe the boundary conditions of the scientific claim, such as potential confounders. Statistical error control is thus only one step in the management of error and uncertainty in theory testing Mayo (2018), Mayo (1996), and Spanos and Mayo (2015). Accepting a theoretical proposition requires corroboration through a variety of methods to ensure robustness and validity, among other concerns. Corroborating evidence is akin (for the sake of analogy) to the concept of justification beyond reasonable doubt in the legal context. It is hard to see how the thesis of epistemic insufficiency could be relevant here, because corroborating evidence is obtained for the most part deductively, via elimination of the relevant alternatives (alternative explanations of particular data patterns or of the total evidence). Although people have questioned the purely deductive status of corroboration, it is at least clear that it is not sufficiently similar to the kind of inductive confirmation that Rudner and Douglas talk about. As these authors also admit, the statistical error rates are not the sole criterion of evidential sufficiency, but it is relatively easier to discuss the fundamental problems of evidential sufficiency using statistical error rates. So, for the sake of argument let's carry on with statistical significance. In §3.4 we will return to this in the context of higher-order uncertainties.

3. Arbitrariness and evidential sufficiency judgments

3.1. What exactly is arbitrary?

The counter-thesis of the inductive risk argument for incorporating values into scientific inference is that scientists should manage inductive risk only with appeal to epistemic considerations.[7] Douglas (2017) and Douglas (2021) claims that this is not possible, because epistemic considerations do not determine why one standard of evidence should be preferred rather than the other–they help determine only the strength of evidence, not whether it is sufficient to accept a hypothesis. Thus, a standard of evidence cannot be set in a non-arbitrary way (unless we take social values on board).

The first question to address is, what exactly is the nature of the charge of arbitrariness? In one intuitive sense of arbitrariness, one can easily avoid it simply by pre-specifying the evidential sufficiency criterion. Imagine a dart shooter who claims to be a sharpshooter,

[5] Not all statisticians agree with the practice of producing dichotomous claims on the basis of evidential thresholds. People who disagree with the use of thresholds also do not need to engage with the inductive risk argument, as they disagree that scientists are in the business of accepting or rejecting scientific hypotheses. Most prominently, this is Jeffrey's response to the inductive risk argument (Jeffrey, 1956). That being said, the majority of scientists across various disciplines use evidential thresholds to make dichotomous claims about phenomena.

[6] A statistical significance level arguably enables us to be able to falsify scientific claims with observations that are known to contain random error. Because the error associated with these observations is probabilistic, such error-containing observations can only be described in probabilistic terms. Since probabilistic statements (e.g., 'This is probably a black swan') by themselves do not have truth values, they are not proper "tests" for scientific claims (i.e., universal statements). Thus, scientists must set an evidential threshold by estimating a confidence interval, which allows them to formulate observation statements that can have truth values and potentially falsify scientific claims (see 'quasi-basic statements', (Uygun Tunç et al., 2023)). This has to do with what Popper calls a methodological choice.

[7] Douglas (2009) attributes this position to Levi (1960).

throws 10 darts and hits the bullseye 6 times, and then defines sharpshooting ability as hitting the bullseye at least 6 times out of 10. Let us also assume that he specifies a different success criterion at each bar, depending on his actual score. This kind of arbitrariness is prevented by specifying any success criteria before the challenge starts. The same argument can be made with respect to the statistical significance level, but scientists already do prevent such arbitrariness when they use the standard error rates in their field (e.g., $\alpha = 0.05$ or 0.00003). The counterpoint, then, presumably is that the standard itself is arbitrary. Elliott (2022) says, "from a purely epistemic perspective, scientists could just as well set a 90 percent or a 99 percent statistical significance level as opposed to a 95 percent statistical significance level."

A slightly more serious charge of arbitrariness thus says that the variability of evidential thresholds testifies to their arbitrariness. The charge of arbitrariness in this case would require the selective application of epistemic standards in determining how much evidential support is adequate. This could indeed be the case if each individual hypothesis was tested using a different evidential threshold, determined in accordance with a context-dependent weighing of inductive risks–just as Douglas recommends. On the contrary, the common practice in science is to use shared standards to prevent judgments of evidential sufficiency from being context-dependent. Furthermore, it should also be noted that the established disciplinary standards regarding the statistical significance levels particularly in social sciences provide additional evidence that social values play no veritable role in determining the evidential thresholds, as studies with wildly different value outlooks use the same α levels for making scientific claims. That is, it is hard to square the claim that evidential sufficiency thresholds necessarily involve social values, with the fact that in disciplines such as economics, psychology, and political science hypotheses with conflicting social value undertones (e.g., liberal, conservative, feminist, male supremacist etc.) are subjected to the same evidentiary standards such as $\alpha = 0.05$.

Wilholt (2009) and Douglas (2017) say that conventional evidential thresholds may solve part of the problem, but those conventions vary across fields. The point is then that the plurality of error thresholds across fields testifies to their arbitrariness. While statistical sets may use some standards of evidential sufficiency in some areas, there is no way to answer this question uniformly "across all judgments" or "across all fields" (Douglas, 2017) . Douglas (2017) opines that different fields thus see different trade-offs between false positives and false negatives as acceptable, where the judgment of acceptability reflects "external concerns" with the use of the knowledge they generate.

This analysis unfortunately displays inadequate understanding of Neyman-Pearson approach to statistical inference and how it is actually used by practicing scientists in testing theories. The specific values at which different fields set their error thresholds vary in accordance with the parameters that go into the calculation of error rates, such as typical sample sizes, the projected base rate of true vs false hypotheses etc. Since these drastically differ from, say, particle physics to psychology or economics, the statistical significance levels also drastically differ (the 5-sigma vs the 1.96-sigma cut-off). It is not the specific evidential thresholds but the error estimation functions that are the same across all fields. Thus, the fact that different fields use different error thresholds testifies not for but against their arbitrariness. It might also be important to note that, when scientists dispute the specific evidential thresholds in their field, they do so on the ground of the epistemic criteria, such as rebalancing the discovery and accuracy trade-off and updating the values that should be assigned to the variables (e.g., prior odds or base rates of true vs false hypotheses) in error estimation functions (see Benjamin et al., 2018). Establishing that statistical significance levels are not arbitrary in any of these senses, we come to the crux of the problem, which the proponents of the inductive risk argument actually do not address, but we will nonetheless.

3.2. A threshold problem for scientific knowledge?

The inductive risk argument maintains that the evidential sufficiency thresholds are necessarily arbitrary. Beyond the options we eliminated, here are two more possible interpretations of arbitrariness in this context. First, the arbitrariness charge can be interpreted as one of precision: whether we can identify one specific value that legitimately distinguishes statistical significance from non-significance. Second, it can be interpreted as a problem of approximation: whether we are able to roughly estimate where that value should be, even if the particular value could never be known or adequately justified.

3.2.1.Arbitrariness as a problem of precision. The thesis of epistemic insufficiency can be considered an analogue of the threshold problem for fallibilism if one interprets it in the precision sense of arbitrary evidential thresholds. Infallibilism about knowledge says that "one can know that p only if one's evidence entails that p," which means that the "conditional probability of p on one's evidence is 1" (J. A. Brown, 2018, p.6). Infallibilism clearly invalidates most (if not all) scientific knowledge. Fallibilism, in rejecting these conditions, is

presumed to hold a threshold view according to which "one can know that p only if its probability on one's evidence exceeds some threshold, t, where that probability threshold is sufficiently high but less than 1" (ibid.). The threshold problem for fallibilism consequently asks what can be given as a "sort of basis or rationale. . . for fixing this level of justification in a non-arbitrary way" (Hetherington, 2001; also see Hannon, 2017).

In the context of scientific inference, let us assume that we accept the Type 1 error of 0.05 to be our evidential threshold as many practicing scientists do. Is not it reasonable that "God loves the .06 nearly as much as the .05"? (Rosnow & Rosenthal, 1992). Even if that may not be so, clearly God would like 0.05001 as much as 0.04999. How can this mark the difference between knowing and not-knowing? The fear of arbitrariness even drives some statisticians to strongly advocate for the practice of interpreting test statistics as continuous measures of discrepancy with a model or to compute continuous measures of evidence (Gibson, 2021; Greenland & Poole, 2013; Rozeboom, 1960; Wasserstein et al., 2019).[8] If the challenge is whether we can justify a single value where the evidential thresholds must be set on purely epistemological grounds, we notoriously cannot (Gigerenzer, 2004; McShane et al., 2019; Rosnow & Rosenthal, 1992; Rozeboom, 1960).

The prospects might look bleak for acquiring scientific knowledge by means of statistical inference. From a purist perspective the issue is obvious: one cannot be sufficiently justified to reject hypothesis H1 at 0.05001 while accepting a similar hypothesis H2 at 0.04999, since they might easily have roughly the same truth value. Hence the impurist solution.

However, once we take this route the problem does not become any less difficult. Let's say we shifted the conventional α-level from 0.05 to 0.09 for studies investigating whether a certain drug increases the women's risk for breast cancer as a side effect, reasoning in light of feminist values that it is more serious an error if the studies fail to find an increase in the study participants when the drug indeed increases the risk, than falsely concluding an increase when the drug is indeed safe. How can we justify the difference between 0.0899 and 0.0901 as the delineation between knowing that the drug is carcinogenic and not-knowing that the drug is carcinogenic using feminist values? Not only we failed to solve the threshold problem, we inflated the false positive risk to the effect that we are less able to know the drug's carcinogenic potential, and less justified in taking any action regarding it. Posing the problem of arbitrariness as one of precision undermines the fallibilist project for anyone, purist or impurist.

The reason is that in the fallibilist framework "is justified" and thereby "knows" are vague predicates. Therefore, they are subject to the Sorites (heap) paradox. The original formulation of the paradox is a modus ponens argument with the following premises (Williamson, 1994):

- 1 grain is not a heap of sand,
- Adding a single grain to a non-heap entity does not make it a heap

Following these premises one should reach the uncontroversial conclusion that 2 grains of sand is not a heap. However, with the repeated application of the same syllogism (i.e., if N grains of sand is not a heap, N+1 grains of sand is also not a heap) one reaches an obviously incorrect conclusion that however big is the number of the grains, there can never be a heap, hence the paradox.

There are a substantial number of solutions suggested to the Sorites paradox (Hyde & Raffman, 2017), but we are not going to delve into these here. The crucial point is that existence of a continuum between two states does not imply the falsity of a distinction between them; otherwise we would not be able to use any 'vague' concept such as young, old, tall, or short meaningfully. Statistical significance, as it is used in common scientific reasoning, can similarly be considered a vague concept, meaning that the borderline between significance and non-significance is somewhat fuzzy. However, the charge of arbitrariness might just lose its claws if we cease to insist on the problem of *precision*.

3.2.2. Arbitrariness as a problem of approximation. Setting the right statistical significance level can also be understood as a problem of approximation. The difficulty of identifying one in practice does not mean that a *real* threshold is not conceivable; one corresponds to an epistemically *optimum* (i.e., maximally reliable) value given the statistical testing theory and the parameters that define the cognitive constraints of the scientific field. Scientists want to have the *maximum number of true positives* and *the minimum number of false negatives* they can achieve in the long run given these constraints. While they may not always be successful in probing into this optimum value, their methodological decision regarding α- and β-levels follows an *algorithmic* optimization of *discovery* vs *accuracy* – two intrinsically scientific, or internal aims. Missing the mark slightly by choosing a conventional threshold that is not identical to the optimum value does not make a fatal difference

[8] As we discussed elsewhere, they fail to see that the function of dichotomous statistical inference is epistemological (i.e., producing testable observation statements that can be used to falsify theoretical claims) rather than statistical, and it is not easily replaceable by alternative approaches (Uygun Tunç et al., 2023)

as long as the conventional thresholds are sufficiently close to the true optimum.

3.3. The case for shared standards

There is another strong rationale for using a *shared* standard, especially one that is not relative to individual hypotheses. That is, scientists want to be able to informatively compare studies and aggregate knowledge. However, this is exactly what the proponents of the inductive risk argument want to abolish when they suggest *customized* or *shifting* evidential thresholds. We can give two reasons for this; an epistemological one having to do with the logic of comparing and combining findings from different studies, and a cognitive one having to do with scientific communication and division of cognitive labor.

Imagine two scientific hypotheses A and B which we would like to test with corrigible data. At a statistical significance level of 0.01, A is accepted and B rejected. At a statistical significance level of 0.05, both A and B are accepted. If A is accepted at the level of 0.01, and B is accepted at that of 0.05, by treating both as evidentially corroborated we violate the logic of knowledge accumulation as long as we do not justify the difference in terms of epistemic criteria such as the base rates. This is not only a matter of scientific communication. Even a solitary agent like Robinson Crusoe, if set out to do a scientific investigation, would need to make a methodological choice to set a threshold to be able to compare a study he did in t1 vs t2.

Secondly, a statistical significance of 0.05 seems to be "low enough such that peers take any claims made with this error rate seriously, while at the same time being high enough such that peers will be motivated to perform an independent replication study to increase or decrease our confidence in the claim" (Uygun Tunç et al., 2023). Clearly lowering the evidential threshold leads to more probative hypothesis tests and thus less uncertainty. But this is meaningful only if there are strong theories in the field which help identify and eliminate systematic sources or error, so that researchers focus on reducing the probabilities of random error. In a field like psychology where there is not yet a base of rich background knowledge to control all causal factors relevant to a phenomenon of interest, gathering huge samples to lower error possibilities (as a field like high energy physics can do) does not make much sense. In this kind of situation, facilitating a faster and more effective process of "conjectures and refutations" serves to create a better division of cognitive labor.

3.4. Higher-order uncertainties

At this point one could argue that the argument developed so far only applies to first-order uncertainties (i.e., uncertainties pertaining to hypotheses). Some have argued that higher-order uncertainties (i.e., uncertainties about uncertainties) are not as tameable as the first-order uncertainties because of a vicious regress problem (i.e., infinite or circular), and that higher-order uncertainties are inevitably non-negligible (M. J. Brown, 2024; Douglas, 2009; Steel, 2016). Although we agree that adequately addressing first-order uncertainties does not always imply that higher-order uncertainties are also adequately addressed, we think both the inevitability of non-negligible higher-order uncertainties and the alleged support this would lend to the inductive risk argument can be contested. This is because 1) higher-order uncertainties are unlikely to involve vicious regress (Henschen, 2021, cf.), 2) they can be managed via systematic higher-order error probing, and 3) scientists cannot be held accountable for inductive risks that stem from yet unknown sources of error.

Our treatment of higher-order uncertainties tracks three possible types of error sources; random, systematic, and unknown. It has been previously shown that the first type (i.e., higher-order uncertainties due to random error) can be handled quite well with standard experimental and statistical procedures (Henschen, 2021). The infinite regress argument suggests that instead of accepting the hypothesis that P(H0)=P0, one would need to assign a probability P1 to this statement, then another probability P2 to P(P(H0)=P0)=P1, and so on, indefinitely. In a well-conducted study, where the statistical distribution assumptions are not violated, there would be quite precise upper and lower bounds of probabilities regarding the first-order uncertainty of random error and the regress is finite because the probability assignments stop at a reasonable point without requiring further probability layers (Henschen, 2021). While occasional instances of circular reasoning may occur, they do not form a general methodological problem, because usually experimental design elements are justified by using criteria other than the experimental outcomes (Henschen, 2021).

Systematic sources of error concern biases or flaws that consistently skew results, such as measurement bias, calibration errors, extraneous factors, observer bias, or selection bias. Concerns with these kinds of issues are typically handled with higher-order error-probing methods in the experimenter's toolset (see, e.g., Galison, 1987; Mayo, 1996). A significant portion of known sources of error are avoided by means of ex ante measures built into experimental planning and design, such as control groups, blinding and proper randomiza-

tion. Others are directly or indirectly probed by means of post hoc measures such as methodological triangulation, robustness checks, replication studies, or meta-analyses. Since statistical tests are commonly not capable of assessing or controlling these kinds or errors, scientists may not conclude their absence on the sole basis of observing a statistically significant effect. Thus, the acceptance of a statistical hypothesis (as decided by a test procedure with a pre-set significance threshold) cannot and does not directly translate into the acceptance of the substantive hypothesis or the scientific claim it is used to test. The justification of statistical significance thresholds themselves thus have little to do (if any) with higher-order uncertainties of this kind, as this is categorically not a problem that can be addressed by adjusting acceptable error rates or any other algorithmic decision protocol (value-free or not). Higher-order error probing is characteristically a deliberative process that requires extensive "experimental knowledge" (Mayo, 1996) and good scientific judgment, thus it typically takes the form of a piecemeal and collaborative (sometimes interdisciplinary) questioning involving multiple researchers and possibly specializations.[9] Often enough, Duhemian kind of underdetermination problems are preemted or resolved. When not, ideally they are not ignored or concealed, which would constitute a breach of epistemic responsibility well before that of a social one.

Lastly, there may of course also be unknown sources of error, which require novel error-probing techniques, scientific discoveries or theoretical advances to be properly identified or controlled. But this is a feature, rather than a bug of ampliative reasoning that is characteristic of science. Since we do not know if there are unknown sources of error, it is very difficult to conceive how inductive risks that may arise due to these can feature in an evaluation of epistemic or other responsibilities (beyond the epistemic responsibility not to assert their absence).

We can thus say that a threshold problem for scientific knowledge does not emerge at the level of higher-order uncertainties either. We last turn to a remaining issue which has not been fully addressed by the preceding.

4. Fallibilism, epistemic value-judgments, and rational disagreement in science

The fallibility of scientific justification implies, as an inherent feature, that epistemic value-judgments have to be made in determining if evidential justification is sufficient for affirming a scientific proposition. By virtue of being value-judgments, these are contestable. This is particularly true when they imply conflicting conclusions about the phenomenon of interest and when it is ambiguous how to weigh different epistemic values in evaluating the existing evidence (Kuhn, 2003). One way to demonstrate this is through the thesis of the theory-ladenness of observation, which says that the evidential criteria for the evaluation of scientific theories are not completely independent from the contents of the very theories they are used to evaluate (Kuhn, 1970). Theory-ladenness implies that different theories might indicate different parts of the existing body of evidence as more important due to the particular set of epistemic values they manifest (more or less strongly than other theories). Since each theory would characterize and assess evidential support partly in reference to its own content, evidential sufficiency decisions are in principle rationally contestable.

However, rationally contestable does not mean arbitrary in any sense that could lend support to the inductive risk argument. First of all, when there is such rational disagreement in a discipline, scientists defer or conditionalize their judgments (Betz, 2013; Giere, 2003; Nagel, 1961). The deferral strategy is the usual first line response. This is because the rational disagreement we outline here is about the differential weighing of empirical evidence in light of epistemic values (with regards to salience or importance), and the effect of divergent epistemic values is inversely proportional to the accumulated evidence (Duhem, 1954, see also). As observations accumulate and more severe tests of the contesting theories are devised, it becomes more and more likely for evidence to overwhelmingly support a theory over its alternatives with little room for doubt, no matter which particular sets of epistemic values are upheld by contesting theories, as many examples of paradigm change demonstrate in the history of science.

Secondly, epistemic values are defined with reference to their truth-conduciveness (Steel, 2010, direct or indirect, see) and thus they are open to empirical evaluation, even if only retrospectively. That is, even if we cannot know at the moment which set of epistemic values is more truth-conducive than the other in a given context, we can reach a justified conclusion about this question over time through increased observation. Hence, even if they seem arbitrary to us at the moment, a retrospective evaluation can determine which set of epistemic values leads to more progress in the Lakatosian sense in the context of a given scientific question. Therefore, the charge of arbitrariness, if it is to be accepted at all, is but rather a temporary issue and strongly tied to not

[9] For an illustration of how such a procedure might work even in the absence of extensive experimental and theoretical knowledge, see our systematic replications framework, (Uygun Tunç & Tunç, 2023)

having enough knowledge about a phenomenon.

Furthermore, based on the successful episodes of theory selection in the history of science, it can be argued that *at least in some cases* rational disagreement is evidently temporary, i.e., it is probable for rational disagreement to evolve into rational consensus as a result of mutual critical engagement and additional evidence. This evolutionary process could be interrupted by the use of social values in scientific inference, which would eliminate (or at least reduce) the need for seeking critical engagement and additional evidence, and thus might lead to premature conclusions. Since it is only retrospectively possible to determine which cases of rational disagreement will evolve into rational consensus (Lakatos, 1978), it may be argued that using social values to resolve theory choice, in response to practical concerns here and now, will be detrimental to the production of scientific knowledge in the long run by stunning the growth of *at least some theories*. One could therefore argue, in principle, for the exclusion of social values from scientific inference, even without believing that all rational disagreements necessarily give way to a consensus.[10]

The first premise of the inductive risk argument, the epistemic insufficiency thesis, seems to be misguided in light of the preceding considerations. The argument requires the value-free ideal either to be successfully defended on an infallibilist basis, or to somehow provide the tools to avoid the logical implications of fallibilism while still committing to it. This amounts to an unsuccessful attempt at *reductuo*, rather than a fair critique. However, even if the reader would disagree with the authors of this paper on the truth of the epistemic insufficiency thesis, there are equally serious problems with the second premise, the thesis of legitimate value-encroachment, which we turn to next.

5. Problems with using non-epistemic values as evidential sufficiency criteria

The critics of the value-free ideal do not necessarily deny the ultimate scientific values that value-freedom is supposed to facilitate, such as veracity. They thus admit that values may have legitimate and illegitimate uses in science. At the core of Douglas (2000, 2009)' demarcation strategy between the legitimate and illegitimate inclusion of non-epistemic values into science lies the distinction between the direct and indirect roles values can play in the context of acceptance. Values play a direct and illegitimate role if they constitute reasons to accept a scientific claim, thereby playing the same role evidence plays. However, they play an indirect and legitimate role if they are used to determine how strong the evidence (or how low the error probabilities) should be to accept a claim. So, according to the proponents of the inductive risk argument, the social legitimacy thesis is valid, since evidence is still the final arbiter of any scientific inference that uses this strategy.

This strategy encounters a distinct set of problems. First and foremost, it is not clear what kind of a meta-criterion determines the extent to which non-epistemic values are allowed to influence evidential standards. The problem becomes evident when we realize that if values are allowed to determine evidential standards for the acceptance of hypotheses, they may also be used to arbitrarily increase or decrease those standards. This will cause evidential standards, such as statistical significance thresholds, to lose their intended function to the effect that it becomes either too easy or too difficult to confirm a scientific claim, which will in turn license bad science or more sophisticated forms of science denialism.

If we think in terms of type 1 and type 2 error probabilities, we see that inflating these error rates beyond a level is no different from the situation where non-epistemic values directly influence scientific conclusions. While the exact location of an error control threshold might be subjected to discussion, we know for a fact that as type 1 error rates increase it becomes increasingly trivial to provide evidential support for false hypotheses. Simmons et al. (2011) have shown that one could even find evidence for factually impossible effects such as "participants get younger if they listen to When I'm Sixty-Four by The Beatles instead of Kalimba by Mr. Scruff" if one simply settles for an elevated type 1 error rate. Similarly, one might play into the hands of science denialism just by allowing the type 2 error rate to be elevated to a point where the existing body of evidence is no longer considered as evidence (Steel & Whyte, 2012, for a related criticism, see). In these examples we do not see a direct influence of non-epistemic values on scientific conclusions, but the outcomes are no different from what a direct influence would have produced. Thus, the allegedly indirect role of determining where to set the thresholds for acceptable error rates can easily turn into evidence manufacturing and/or denial.

The question is then, how can the social legitimacy thesis prevent conclusions that are nothing but wishful thinking with sprinkles of technical jargon? Specifi-

[10] In this context, it can be argued that these kinds of problems are more likely to be encountered in younger/less mature disciplines. It can also be argued that there is a serious risk associated with allowing the social values to be used as criteria of evidential sufficiency in such disciplines as this may stunt the epistemic iterative processes (Chang, 2004) that would help the discipline mature its evidence collection methodologies.

cally, *how shall we determine where an indirect role stops and a direct role begins non-arbitrarily without appeal to strictly epistemic criteria*? If we accept the need to introduce an epistemic constraint, then it becomes difficult to say that non-epistemic values indeed play a veritable role. This is because, if the meta-criterion for delineating direct and indirect influences of social values is strictly epistemic in nature, the argument would actually turn into a reformulation of the value-free ideal, instead of its refutation. On the other hand, if we reject the need to constrain how much weight non-epistemic values can have in determining where to set the error control thresholds, then the indirect role evaporates and we are left only with a direct role, which Douglas criticized as being illegitimate. Thus, Douglas' distinction between legitimate and illegitimate uses of social values in science suffers from a *paradox* of meta-criterion.

The downstream consequences of the paradox of meta-criterion become evident when the proposal of setting value-laden evidence thresholds is regarded as a strategy to resolve underdetermined theory choice situations. As a methodological problem, theory choice requires scientists to reduce the risk of misallocating empirical support among rival hypotheses, and the only epistemically rational way to do this is by mitigating the underdetermination of scientific tests. This route often takes time, but terminating it prematurely (by using values in resolving theory choice) not only fails to manage underdetermination but exacerbates it.

We encounter another problem if we opt to apply the same strategy to deal with the underdetermination of scientific tests, namely by using non-epistemic values to weigh the risks of erroneously rejecting or maintaining a hypothesis due to false auxiliary assumptions or a false ceteris paribus clause. Douglas (2000) (p. 565; see also Biddle and Kukla, 2017) famously extends the inductive risk argument to apply to all kinds of scientific decisions preceding the final decision to accept or reject a scientific claim, such as decisions regarding data collection, analysis and interpretation, or the choice of methodology. The problem with applying non-epistemic values to weigh inductive risks is that regardless of whether we focus on the epistemic or practical consequences of error, our risk mitigation strategy should effectively reduce the probability of making false auxiliary assumptions or making a false ceteris paribus assumption. It can be easily seen that values are largely irrelevant in reducing these probabilities. Moreover, if values are allowed in the selection of instruments, models, parameters, outlier data points, or background theories and facts, what prevents these decisions from being ad hoc? If there are no epistemic constraints to prevent the ad hoc selection of auxiliary assumptions, we can easily fall back to the problem of wishful thinking. On the other hand, if we set epistemic (and hence value-neutral) meta-criteria to determine the legitimate extent of social value influences on evidential thresholds, we manifestly circle back to a value-free position, where social values are only allowed to influence the initial preference for weighing some particular epistemic concerns over others, as a subjective decision which is to be discarded in the long run through collection of more persuasive evidence (as suggested by, Kuhn, 2003).

6. Epistemic problems require epistemic answers

Last but not least we must address the question: Evidential sufficiency for which decision? There are indeed not one but two questions:

> 1) Which evidential threshold would optimally manage different kinds of error probabilities (so that the term evidence is meaningfully applicable)?
>
> 2) Which level of evidential readiness would be most optimal to realize objective P in view of the (social, moral, or political) value V?

As it has been indicated previously by others (most notably Levi, 1960), this second question is the one the proponents of the inductive risk argument are indeed asking, but they present it as if this question has to do with the interpretation of evidence. Technically speaking, the theory of statistical inference (or hypothesis testing) and the statistical decision theory (or decision-making under uncertainty) are distinct, despite having common historical origins and featuring the term 'decision' in relation to a situation involving uncertainty. A decision in the context of statistical hypothesis testing (to which Both Rudner and Douglas refer) is *evidential*; it is an aspect of *inference* (Birnbaum, 1977; Levi, 1962; Mayo, 1996). A decision in the statistical decision theory is *behavioral*, hence much closer to the literal or intuitive sense of the word. Birnbaum (1977) compares the two decision schemas to highlight the key differences. In the context of statistical decision theory, a typical example of a decision is "place [a given] batch [of products] in the market" vs "withhold the batch from the market," which are concrete actions in the ordinary sense. In the context of data analysis, a direct application of the same schema may lead one to think that the corresponding actions are "reject H0" vs "do not reject H0," where H0 is the null hypothesis, whereas a more appropriate comparison would characterize the decision options as "reject H0 for H1, α, β" vs"reject H1 for H0, α, β," where H1 is the alternative hypothesis, and α and β are the error probabilities. Using ordinary semantical

formulations one could express the difference as that between 'deciding *that* evidence E corroborates hypothesis H' and 'deciding *to* act in a certain way that is in accordance with the truth of H' (cf. Birnbaum, 1977). Cox (1958, p.354) similarly characterizes the first as "deciding what types of statements can usefully be made and exactly what they mean," whereas "in statistical decision theory. . . the possible decisions are considered as already specified."

The meaning of 'acceptance' in reference to statistical hypotheses in science corresponds to the inferential sense of decision and not the behavioral. In Levi (1960)'s analysis, Rudner (1953)'s conclusion–that evidential sufficiency judgments should reflect the importance of error–depends on an implicit premise: "To choose to accept a hypothesis H as true (or to believe that H is true) is equivalent to choosing to act on the basis of H relative to some specific objective P." Without the *pragmatist* move to identify the two, the inductive risk argument is invalid.

One might allude to the fact that the most widely accepted conceptualization of statistical hypothesis testing, the behavioral account developed by Neyman (1950, pp.258-259), seems to be compatible with a pragmatist identification of the above kind. However, similar to the term 'decision', 'behavior' also takes on a special, epistemic meaning in scientific statistical inference. As we explained elsewhere, the most plausible way to understand the meaning of behavior in the Neyman-Pearson approach to scientific hypothesis testing is that of considering an observation as expressing a "singular fact" and acting on the basis of it in an ongoing process of inquiry (Uygun Tunç et al., 2023). Acting on the basis of H in scientific research means using it as background knowledge to test a subsequent hypothesis, where the relevant "objective" is the growth of knowledge through further inquiry–an epistemic one. Therefore, even if we accept Rudner's implicit premise, the question according to which the long term error rates should be determined in science is still the first question (i.e. Which evidential threshold would optimally manage different kinds of error probabilities so that the term evidence is meaningfully applicable?) but not the second.

Another probable source of confusion is that inferential or evidential decisions are not merely about cognitive attitudes or theoretical/methodological commitments in an ongoing process of inquiry but also involve sharing one's conclusions with a wider community of fellow scientists. From one perspective, it is where the line that demarcates the evidential and behavioral type of decisions is blurred, because publishing one's findings is an action with potential ethical responsibilities. Granting this much, we still have a problematic conflation at our hands, that goes beyond the debate on what decisions mean in statistical inference. The conflation concerns essentially different kinds of action, namely *speech* and *conduct*. The categories of speech act and conduct are distinct, and have very diverse implications in terms of ethical responsibility. A scientific report, as a kind of *assertion*, must be honest and accountable, namely it must disclose and justify the methods and standards adopted to collect, analyze, and interpret the data, and should not overclaim, i.e., assert conclusions the research procedure does not license. Scientifically informed policy, as *conduct*, must assess and evaluate the ethical, economic, social and possibly political stakes involved, and be in a position to explain and defend the methods of assessment and the principles evaluation. This important distinction is completely neglected even in the more recent formulations of the inductive risk argument (see, Havstad, 2022)[11]. The responsibility assigned to scientists by the inductive risk argument is one of *conduct*, yet publishing your findings is a *speech act* par excellence, therefore the ethical responsibilities must be judged accordingly. If the proponents of the inductive risk argument reject any relevant distinction between the category of speech acts and that of conduct, a controversial move if not inadmissible, they must explicitly and successfully argue for this.

Second, when we have a closer look at the actual content of speech in scientific publications, we overwhelmingly see linguistic "stance markers" which convey some degree of uncertainty or conditionality even if the scientific claims are not explicitly hedged [12]. Using verbs such as 'suggests' instead of 'demonstrates' or 'proves', or adding sentential adverbials such as 'as far as we know' is very typical of scientific jargon, especially for causal claims or idealized models, because of the inconclusiveness of scientific justification or the fact that most scientific knowledge has boundary conditions in its application. To the extent that a scientific report transparently and honestly discloses the uncertainties associated with the claims, as a speech act it cannot be taken as an assurance (which signals certainty), and it is at least questionable that it can even be taken as an

[11] This particular formulation is argued to be valid and sound by some scholars (M. J. Brown, 2024; M. J. Brown & Stegenga, 2023, see). However, because the premises 1-4 fail to distinguish different types of decisions (i.e., inferential and behavioral) and actions (i.e., speech and conduct) they are arguably false, and thus the recent formulation can be said to be valid but not sound.

[12] See also Benton and Van Elswyk (2020) on hedged assertions.

assertion (which signals knowledge)[13]. Consequently, it is very difficult to see how such a report generates a responsibility in reference to the risks associated with the conduct that assumes the truth of what has been said and commits to act on it.

But nonetheless, the invalidity of the inductive risk argument does not completely hang on the distinction between two kinds of decisions or actions, that is between belief and action or speech and conduct. Even if we understood both decisions in a somewhat similar sense, the objectives are completely different in nature: Any objectives that may feature in setting the evidential standards for scientific inference are 'theoretical' or 'epistemic'. Due to the difference in the nature of the objectives the second kind of decision is typically insulated from the first; *the stakes influencing it do not encroach on the first kind of decision context*. For instance, policy makers sometimes decide to act on the basis of weak scientific evidence for social considerations, which does not feed back into science - no scientist endorses or even pursues a scientific proposition because it is used as a basis for some societal intervention.

We can speculate that the reverse is also true: The stakes that might influence epistemic decisions do not encroach on the pragmatic decision contexts. The proponents of the pragmatic encroachment thesis derive their motivation from the epistemic norms of practical rationality. Hookway (1990, p.139)says about justification that "our understanding of the amount of evidence we require in support of an hypothesis before we can describe it as justified may reflect the degree of support that is required before we can feel that we are acting responsibly when we act upon it." Similarly, Fantl and McGrath (2002, p. 78) say that "S is justified in believing that p only if S is rational to act as if p." However, engineers use Newtonian mechanics in all kinds of applications all the while knowing that it is evidently false, and do not prefer to use the better corroborated alternatives since these do not increase but decrease practical utility. Hence, this seems to be an odd way to specify necessary conditions for knowledge and action. Practical action has many norms which may be in conflict or tension with one another. The epistemic norms of action (such as acting on the basis of knowledge or justified belief) are not immune to being overridden by moral norms, for instance if not acting due to insufficient evidence would create certain harm. In the case of scientific policy advice, the policy makers might rationally decide, given various factual and normative considerations, to act on the basis of merely plausible opinions by the scientific advisers, or they may even rationally decide not to accept a very narrow error margin that satisfies the scientific researchers as practically unacceptable.

Importantly, if the pragmatic considerations affecting the decision making context are allowed to alter the conditions of epistemic justification, which are then used to justify practical action per the epistemic norms of practical rationality, we end up with a vicious type of circularity. To illustrate, in the error-theoretical framework error control serves risk management, which may also involve societal risks. If risk-driven concerns are allowed to influence the decisions pertaining to error control, which Douglas' argument implies, error control may not properly serve risk management. This meta-risk will arise especially in contexts where epistemic and non-epistemic values can compete. The balancing of the type 1 and type 2 error probabilities employs epistemic values, where scientists try to optimize the trade-off between discovery and accuracy for a given research domain. When non-epistemic values are allowed to compete with epistemic values, they will necessarily divert from this epistemic optimum. In the long run, scientifically informed policy will be less effective if risk management guides error control rather than being guided by it. This in turn will be detrimental to the very non-epistemic values that inform risk management judgments. There are important reasons why we should answer epistemic questions with epistemic answers. Otherwise, we come up with euphemisms for bias.

We argued that the inductive risk argument fails to justify the key premise of epistemic insufficiency beyond what is entailed by fallibilism about scientific knowledge, and its second premise of legitimate value-encroachment suffers a meta-criterion problem. In closing the paper, let's cast the net a bit wider. However unlikely, let's imagine that a value or set of values are proposed with the promise of avoiding the problem of meta-criterion, on the grounds that while themselves not being epistemic values, they are not epistemically arbitrary either. This approach would not solve but create yet another meta-criterion problem, as we argue next.

[13] Some authors have suggested that scientific publications violate or do not conform to the ordinary norms of assertion, most commonly, the knowledge norm. For instance, Dang and Bright (2021) argue that scientific publications require neither belief nor truth nor justification. Dethier (2022) argues, on the other hand, that some scientific assertions can still be properly made without being known or justifiably believed: A scientific proposition can be "advanced" in a publication, by saying, for instance, that "The evidence provided by the present study supports P."

7. The problem of meta-criterion reappears

Some philosophers take a cognitivist/realist perspective on social/moral values and reject the primacy of epistemic values over social values in scientific inference (M. J. Brown, 2013, 2017). According to this approach, some social values may be based on better reasons than others, and to the extent that they are based on better reasons (rather than subjective preferences), they can be used as inference criteria like epistemic values (M. J. Brown, 2017). Good reasons can be empirical as well as moral, since, according to these scholars, there is a two-way relationship between social values and empirical theories (Anderson, 2004; Nelson, 1990). Defined as such, right values are not mere preferences (thus not associated with wishful thinking) and squarely belong to any sound scientific reasoning.

The proponents of the right values approach arguably attribute to social values the uncertainty-reducing functions previously attributed to epistemic values, such as determining the auxiliary assumptions in a way that would increase our ability to evaluate competing theories (Kuhn, 2003; Laudan, 1978). Since epistemic values are defined in terms of being truth-conducive (Laudan, 1984; McMullin, 1982; also see Steel, 2010 for intrinsic and extrinsic distinction), their uncertainty-reducing function is relatively uncontroversial, at least in the long run. Social values by definition have no intrinsic property of truth-conduciveness, irrespective of whether they are “right” or not. Even assuming that right values interact with facts, it is not clear why it would be better to take the indirect route of appealing to right values in epistemic inference rather than directly using the data and epistemic criteria that have shaped these values. It is therefore questionable whether social values, even if they are "right," would constitute the right *reasons* in the context of scientific inference.

Furthermore, the uncertainty-reducing function of right values depends on these systems of values leading to consistent, or at least converging results when applied to different test situations. Since there is a clear meta-criterion for epistemic values (i.e., truth/verisimilitude) it can be argued that utilizing epistemic values would result in increasingly consistent results, at least in the long run. In terms of social values, there is no such clear meta-criterion, and finding one may not be possible. Also, even admitted by other critics of the value-free ideal, more often than not the implications of social values for the practical scientific inquiry are very vague and it is very hard for scientists to conceptualize and anticipate such implications in a sufficient manner (de Melo-Martín & Intemann, 2016).

This problem concerns the existence of a value-independent meta-criterion for the selection of right values. In the absence of a strictly value-independent meta-criterion, we encounter a circularity issue, that is, the argument for the right values would still be dependent on the same or associated values. The issue of circularity with respect to the choice of right values can also be conceptualized as in the Mannheim Paradox, which was previously formulated in the context of political ideologies. The Mannheim Paradox points to the impossibility of a completely ideology-free perspective when evaluating ideologies (Breiner, 2013; Geertz, 2014; Ricœur, 1986). Since the criteria used to evaluate the ideologies are themselves necessarily (at least partially) ideological, it is impossible to have a completely non-ideological position on ideologies. In other words, the proposition that a value system is the "right" value system cannot be justified independent of criteria that are associated with that same value system, and hence will be necessarily circular.

The circularity issue described here should not be thought of as a purely logical problem, because it would also possibly lead to a coordination issue. Ignoring for the moment the conundrum that it seems impossible to find a *value-independent yet non-epistemic meta-criterion* for choosing the right values, there is also the risk that the values of the social groups that have an advantage in the process of determining the “right” values will be absolutized in the garb of scientific facts. Even a meta-criterion such as “harm-avoidance” (see Douglas, 2009), on which one might assume that there is a broad consensus, can be argued to involve some ideological contestation. Previous studies show that harm avoidance is of different degrees of importance to liberal and conservative individuals, and that conservatives are more likely than liberals to use criteria other than harm in their value judgments (Kivikangas et al., 2021). That is, in any value trade-off situation liberals tend to weigh harm-avoidance as more important than other values compared to conservatives (Graham et al., 2009). From this perspective one might argue that the adoption of harm-avoidance as a meta-criterion reflects the predominantly liberal worldview of academics, especially in social sciences and humanities (Cardiff & Klein, 2005).

8. Conclusion

In this paper, we have investigated the two key premises of the inductive risk argument against the value-free ideal of science, the theses of epistemic insufficiency and legitimate value-encroachment. Our analysis, particularly regarding the purported arbitrariness of evidential thresholds, showed that the inductive risk argument does not demonstrate the untenability of the value-free ideal of science beyond what can already be

inferred from the thesis of fallibilism regarding scientific knowledge. Since the untenability of the value-free ideal cannot be deduced solely from fallibilism about knowledge, the inductive risk argument falls short of justifying the epistemic insufficiency thesis. Our analysis also indicates that incorporating social values into judgments of evidential sufficiency would *weaken* the justification of scientific inferences due to lack of a meta-criterion for the legitimate use of social values in scientific inference. The problem of meta-criterion leads to one or more major problems, depending on the argumentative strategy adopted; namely wishful thinking (due to failure to indicate a meaningful distinction between epistemically legitimate and illegitimate uses of social values), category mistakes in identifying the relevant parameters for decision-making (due to failure to identify the actual stakes involved in epistemic vs practical domains), and the Mannheim-style paradoxes of social legitimacy (due to failure to justify social values in a non-circular manner). Thus, the thesis of legitimate value-encroachment is wrong, as value-ladenness *exacerbates* the inferential risks rather than *resolving* or *diminishing* them.

The ball is now in the court of the proponents of the inductive risk argument to show if and how the inductive risk argument 1) poses a distinct epistemic challenge that cannot already be inferred from fallibilism about knowledge and 2) identify a rational method for determining a meta-criterion that would not lead to problems such as wishful thinking, illegitimate encroachment between domains, and the Mannheim Paradox. As it stands, the inductive risk argument does not seem to establish the untenability of the value-free ideal and the need for a value-laden alternative.

Author Contact

Duygu Uygun Tunç / Society of Fellows in the Liberal Arts / University of Chicago, USA ORCID: https://orcid.org/0000-0003-0148-0416

Mehmet Necip Tunç / Social Psychology Department / Tilburg University, the Netherlands ORCID: https://orcid.org/0000-0002-1350-174X

Conflict of Interest and Funding

There are no conflicts of interest associated with this research.

Author Contributions

Both authors have contributed equally to conceptualization, project administration, original draft preparation, review, and editing.

Acknowledgements

The authors would like to thank Katharina Bernhard, Matthew J. Brown, Gregor Gaszczyk, Sandy Goldberg, Gürol Irzık, Daniël Lakens, and Oliver Maclaren for critical comments and suggestions on earlier versions of this paper, which helped to substantially improve it.

References

Anderson, E. (2004). Uses of value judgments in science: A general argument, with lessons from a case study of feminist research on divorce. *Hypatia*, *19*(1), 1–24. https://doi.org/10.1111/j.1527-2001.2004.tb01266.x

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., & Camerer, C. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*(1), 6–10.

Benton, M. A., & Van Elswyk, P. (2020). Hedged assertion. *The Oxford Handbook of Assertion*.

Betz, G. (2013). In defence of the value free ideal. *European Journal for Philosophy of Science*, *3*(2), 207–220.

Biddle, J. B., & Kukla, R. (2017). The geography of epistemic risk. In K. C. Elliott & T. Richards (Eds.), *Exploring inductive risk: Case studies of values in science* (pp. 216–237). Oxford University Press.

Birnbaum, A. (1977). The neyman-pearson theory as decision theory, and as inference theory; with a criticism of the lindley-savage argument for bayesian theory. *Synthese*, *36*(1), 19–49.

Bogen, J., & Woodward, J. (1988). Saving the phenomena. *The Philosophical Review*, *97*(3), 303.

Breiner, P. (2013). Karl mannheim and political ideology. *The Oxford Handbook of Political Ideologies*.

Brown, J. A. (2018). Fallibilism: Evidence and knowledge. *Oxford University Press*.

Brown, M. J. (2013). Values in science beyond underdetermination and inductive risk. *Philosophy of Science*, *80*(5), 829–839.

Brown, M. J. (2017). Values in science: Against epistemic priority. *Current Controversies in Values and Science*.

Brown, M. J. (2024). For values in science: Assessing recent arguments for the ideal of value-free science. *Synthese*, *204*(4), 112.

Brown, M. J., & Stegenga, J. (2023). The validity of the argument from inductive risk. *Canadian Journal of Philosophy*, *53*(2), 187–190. https://doi.org/10.1017/can.2023.37

Cardiff, C. F., & Klein, D. B. (2005). Faculty partisan affiliations in all disciplines: A voter-registration study. *Critical Review*, *17*(3-4), 237–255.

Chang, H. (2004). *Inventing temperature: Measurement and scientific progress*. Oxford University Press.

Cox, D. R. (1958). Some problems connected with statistical inference. *The Annals of Mathematical Statistics*, *29*(2), 357–372. https://www.jstor.org/stable/2237334

Dang, H., & Bright, L. K. (2021). Scientific conclusions need not be accurate, justified, or believed by their authors. *Synthese*, *199*(3–4), 8187–8203.

de Melo-Martín, I., & Intemann, K. (2016). The risk of using inductive risk to challenge the value-free ideal. *Philosophy of Science*, *83*(4), 500–520. https://doi.org/10.1086/687933

Dethier, C. (2022). Science, assertion, and the common ground. *Synthese*, *200*(1), 30.

Douglas, H. (2000). Inductive risk and values in science. *Philosophy of Science*, *67*(4), 559–579.

Douglas, H. (2009). *Science, policy, and the value-free ideal*. University of Pittsburgh Press.

Douglas, H. (2017). Why inductive risk requires values in science. In *Current controversies in values and science* (pp. 81–93). Routledge.

Douglas, H. (2021). *The rightful place of science: Science, values, and democracy: The 2016 descartes lectures*. Consortium for Science, Policy & Outcomes.

Duhem, P. (1954). *The aim and structure of physical theory* [Translated by Philip P. Wiener. Original work published 1906]. Princeton University Press.

Elliott, K. C. (2011). Direct and indirect roles for values in science. *Philosophy of Science*, *78*(2), 303–324.

Elliott, K. C. (2022). *Values in science*. Cambridge University Press.

Fantl, J., & McGrath, M. (2002). Evidence, pragmatics, and justification. *The Philosophical Review*, *111*(1), 67–94.

Fantl, J., & McGrath, M. (2009). Advice for fallibilists: Put knowledge to work. *Philosophical Studies*, *142*, 55–66.

Galison, P. L. (1987). *How experiments end*. University of Chicago Press.

Geertz, C. (2014). Ideology as a cultural system. In *Ideology* (pp. 279–294). Routledge.

Gibson, E. W. (2021). The role of p-values in judging the strength of evidence and realistic replication expectations. *Statistics in Biopharmaceutical Research*, *13*(1), 6–18. https://doi.org/10.1080/19466315.2020.1854044

Giere, R. N. (2003). A new program for philosophy of science? *Philosophy of Science*, *70*(1), 15–21.

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, *33*(5), 587–606. https://doi.org/10.1016/j.socec.2004.09.033

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, *96*(5), 1029.

Greenland, S., & Poole, C. (2013). Living with p values: Resurrecting a bayesian perspective on frequentist statistics. *Epidemiology*, *24*(1), 62–68. https://doi.org/10.1097/EDE.0b013e3182785741

Hannon, M. (2017). A solution to knowledge's threshold problem. *Philosophical Studies*, *174*(3), 607–629. https://doi.org/10.1007/s11098-016-0700-9

Havstad, J. C. (2022). Sensational science, archaic hominin genetics, and amplified inductive risk. *Canadian Journal of Philosophy*, *52*(3), 295–320.

Havstad, J. C., & Brown, M. J. (2017). Neutrality, relevance, prescription, and the ipcc. *Public Affairs Quarterly*, *31*(4), 303–324.

Hempel, C. G. (1960). Science and human values. In R. E. Spiller (Ed.), *Social control in a free society* (pp. 39–64). University of Pennsylvania Press.

Henschen, T. (2021). How strong is the argument from inductive risk? *European Journal for Philosophy of Science*, *11*(3), 92.

Hetherington, S. (2001). *Good knowledge, bad knowledge: On two dogmas of epistemology*. Clarendon Press.

Hookway, C. (1990). *Scepticism*. Routledge.

Hyde, D., & Raffman, D. (2017). Sorites paradox. In E. N. Zalta (Ed.), *Stanford encyclopedia of philosophy*. Stanford University. https://plato.stanford.edu/entries/sorites-paradox/

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

Kim, B. (2017). Pragmatic encroachment in epistemology. *Philosophy Compass*, *12*(5), e12415. https://doi.org/10.1111/phc3.12415

Kivikangas, J. M., Fernández-Castilla, B., Järvelä, S., Ravaja, N., & Lönnqvist, J. E. (2021). Moral foundations and political orientation: System-

atic review and meta-analysis. *Psychological Bulletin*, *147*(1), 55.

Kuhn, T. S. (1970). *The structure of scientific revolutions*. University of Chicago.

Kuhn, T. S. (2003). Objectivity, value judgment, and theory choice. In A. Bird & J. Ladyman (Eds.), *Arguing about science*. Routledge.

Lacey, H. (2015). 'holding' and 'endorsing' claims in the course of scientific activities. *Studies in History and Philosophy of Science Part A*, *53*, 89–95.

Lakatos, I. (1978). *The methodology of scientific research programmes* (J. Worrall & G. Currie, Eds.). Cambridge University Press.

Laudan, L. (1978). *Progress and its problems: Towards a theory of scientific growth*. University of California Press.

Laudan, L. (1984). *Science and values: The aims of science and their role in scientific debate*. University of California Press.

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

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

Lusk, G. (2021). Does democracy require value-neutral science? analyzing the legitimacy of scientific information in the political sphere. *Studies in History and Philosophy of Science Part A*, *90*, 102–110.

Mayo, D. G. (2018). *Statistical inference as severe testing: How to get beyond the statistics wars*. Cambridge University Press.

Mayo, D. G. (1996). *Error and the growth of experimental knowledge*. University of Chicago Press.

McMullin, E. (1982). Values in science. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, *1982*(2), 2–28.

McShane, B. B., Gal, D., Gelman, A., Robert, C., & Tackett, J. L. (2019). Abandon statistical significance. *The American Statistician*, *73*(sup1), 235–245. https://doi.org/10.1080/00031305.2018.1527253

Menon, T., & Stegenga, J. (2023). Sisyphean science: Why value freedom is worth pursuing. *European Journal for Philosophy of Science*, *13*(48), 1–24. https://doi.org/10.1007/s13194-023-00552-7

Miller, B. (2024). *The social dimensions of scientific knowledge: Consensus, controversy, and coproduction*. Cambridge University Press.

Nagel, E. (1961). *The structure of science: Problems in the logic of scientific explanation*. Harcourt, Brace, World.

Nelson, L. H. (1990). *Who knows: From quine to a feminist empiricism*. Temple University Press.

Neyman, J. (1950). *First course in probability and statistics*. Henry Holt.

Ricœur, P. (1986). *Lectures on ideology and utopia* (G. H. Taylor, Ed.). Columbia University Press.

Rosnow, R. L., & Rosenthal, R. (1992). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist*.

Rozeboom, S. (1960). The fallacy of the null-hypothesis significance test. *Psychological Bulletin*, *57*(5), 416–428. https://doi.org/10.1037/h0042040

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

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366.

Spanos, A., & Mayo, D. G. (2015). Error statistical modeling and inference: Where methodology meets ontology. *Synthese*, *192*, 3533–3555.

Stanley, J. (2005). Fallibilism and concessive knowledge attributions. *Analysis*, *65*(2), 126–131.

Steel, D. (2010). Epistemic values and the argument from inductive risk. *Philosophy of Science*, *77*(1), 14–34.

Steel, D. (2016). Climate change and second-order uncertainty: Defending a generalized, normative, and structural argument from inductive risk. *Perspectives on Science*, *24*(6), 696–721. https://doi.org/10.1162/POSC_a_00229

Steel, D., & Whyte, K. P. (2012). Environmental justice, values, and scientific expertise. *Kennedy Institute of Ethics Journal*, *22*(2), 163–182. https://doi.org/10.1353/ken.2012.0010

Steele, K. (2012). The scientist qua policy advisor makes value judgments. *Philosophy of Science*, *79*(5), 893–904.

Stegenga, J. (2024). Fast science. *The British Journal for the Philosophy of Science*, 729617.

Stegenga, J., & Menon, T. (2023). The difference-to-inference model for values in science. *Res Philosophica*, *100*(4), 423–447. https://doi.org/10.5840/resphilosophica2023928102

Suppes, P. (1974). The structure of theories and the analysis of data. In F. Suppe (Ed.), *The structure of scientific theories* (pp. 266–283). University of Illinois Press.

Uygun Tunç, D., & Tunç, M. N. (2023). A falsificationist treatment of auxiliary hypotheses in so-

cial and behavioral sciences: Systematic replications framework. *Meta-Psychology*, *7*.

Uygun Tunç, D., Tunç, M. N., & Lakens, D. (2023). The epistemic and pragmatic function of dichotomous claims based on statistical hypothesis tests. *Theory & Psychology*, *33*(3), 403–423.

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond "p < 0.05". *The American Statistician*, *73*(sup1), 1–19. https://doi.org/10.1080/00031305.2019.1583913

Wilholt, T. (2009). Bias and values in scientific research. *Studies in History and Philosophy of Science Part A*, *40*(1), 92–101.

Williamson, T. (1994). *Vagueness*. Routledge.

Woodward, J. (1989). Data and phenomena. *Synthese*, *79*(3), 393–472.

Woodward, J. (2000). Data, phenomena, and reliability. *Philosophy of Science*, *67*(Supplement), S163–S179. https://doi.org/10.1086/392817