On Sober's Criterion of Contrastive Testability

Sebastian Lutz*

Draft: 2010-12-13

Abstract

Elliott Sober has suggested his criterion of contrastive testability as an improvement over previous criteria of empirical significance like falsifiability. I argue that his criterion renders almost any theory empirically significant because its restrictions on auxiliary assumptions are to weak. Even when the criterion is modified to avoid this trivialization, it fails to meet other conditions of adequacy for a criterion of empirical significance that follow from Sober's position. I suggest to define empirical significance as empirical non-equivalence to a tautology, because this definition does meet the conditions of adequacy. Specifically, it is equivalent to the standard Bayesian criterion of empirical significance whenever all probabilities are defined and contains falsifiability as a special case. This latter feature is important because those conditions of adequacy that apply to criteria of deductive empirical significance single out falsifiability.

Contents

*Theoretical Philosophy Unit, Utrecht University. sebastian.lutz@gmx.net. I thank James Justus, Jesse Mulder, and especially Thomas Müller and Elliott Sober for helpful comments and discussions.

1 Introduction

Over the last two decades, Elliott Sober (1990; 1999; 2007; 2008) has developed and defended a criterion of empirical significance, called "testability", that is both promising and much needed. The promise stems from Sober's defense of the criterion's basic assumptions and the fact that it can deal with probabilistic theories and auxiliary assumptions. The need for a criterion of empirical significance stems from, for example, questions about the empirical significance of string theory (Smolin 2006, Woit 2006) and the theory of intelligent design (ID), to which Sober (1999; 2007; 2008) applies his criterion. Given the possible applications of a criterion of empirical significance and the simultaneous wide-spread belief that the search for such a criterion has utterly failed (Soames 2003, ch. 13), it is somewhat surprising that Sober's criterion has neither been subjected to much scrutiny nor led to further research into criteria of empirical significance.[1] This article is meant to fill this gap.

I will only shortly discuss Sober's application of testability, and argue that it cannot be used to show that ID is not empirically significant (§2). More importantly, I will argue that the criterion as it stands is trivial, but can be saved by changing the restrictions on the auxiliary assumptions (§3).

Since Sober considers his definition of confirmation an explication (Sober 2008, 35) and his definition of testability is an outgrowth thereof, it is plausible that the latter is also meant as an explication.[2] His definition is thus an explicatum for the explicandum that could be circumscribed by pre-analytic terms like 'having empirical content', 'making observational assertions', 'predicting experimental outcomes', or just 'being an empirical theory' (cf. Kuipers 2007).[3] One desideratum of an explication is that it must be possible to use the explicatum in place of the explicandum in the relevant contexts (Carnap 1950, §3; Hempel 1952, 663), which often leads to a variety of conditions of adequacy that an explicatum has to fulfill. I will argue that Sober's assumptions and the intended application of his criterion of empirical significance lead to eight conditions of adequacy (§4.1), one of them the condition that a probabilistic criterion of empirical significance should contain falsifiability as a special case (§4.2).

[1] In fact, the only discussion of the criterion itself, rather than its application to ID, might be given by Justus (2010) in a book review.

[2] Sober (2010, 1) states as much in an unpublished note.

[3] I will distinguish between use and mention and between concepts and their names only when this improves clarity or readability.

I then discuss the role of undefined probabilities to elucidate the conditions of adequacy for probabilistic theories (§5) and show that Sober's definition of testability meets only four of the conditions of adequacy (§6). But this article is not purely critical. As a positive contribution to the discussion, I suggest to define that a theory makes observational assertions if and only if it is not empirically equivalent to a tautology, and to call a theory empirically significant if and only if it makes observational assertions. This criterion of empirical significance relies only on concepts that Sober accepts and fulfills all eight conditions of adequacy (§7).

2 Three proposals for a criterion of empirical significance

Sober (2008, 154) conjectures that his criterion is "a step forward from the failed proposals of the logical positivists", but this is misleading because the logical positivists wanted to distinguish between meaningful and meaningless sentences (cf. Carnap 1963, §6.A). Sober (2010, 1), on the other hand, does not consider his definition of testability to provide a criterion of empirical significance precisely because "'[e]mpirical significance' suggests that a sequence of terms has meaning iff it is empirically testable" and he does not subscribe to this position. Rather, Sober (2008, 149f) argues, meaningfulness is a semantic concept, while testability is epistemic. And furthermore:

> It seems clear that meaningfulness and testability are different. I suppose that the sentence "undetectable angels exist" is untestable, but the sentence is not meaningless gibberish. We know what it says, what logical relations it bears to other statements, and we can discuss whether it is knowable; none of this would be possible if the string of words literally made no sense.

Therefore Sober is rather improving on the demarcation criterion by Popper, who wanted to distinguish empirical from non-empirical statements, both of which can be meaningful (Popper 1935, §4; 1963, §II). I will follow Sober in the search for a demarcation criterion for empirical statements, but not in his choice of terminology. I think that 'empirical significance' differs from 'meaningfulness' enough to avoid confusion, and, as a technical term, is clearly meant as a placeholder for an explicatum. In any case, nothing in the following will hinge on this terminology.

In line with his search for a demarcation criterion, Sober (2007; 2008, §2.8) introduces his criterion as avoiding the problems of Popper's falsifiability criterion, which demands "that the theory allow us to deduce, roughly speaking, more empirical singular statements than we can deduce from the initial conditions alone" (Popper 1935, 85). By application of *modus tollens*, and assuming that the negation of an observation sentence is itself an observation sentence, Popper (1935, 86) arrives at the following definition:

> A theory is to be called 'empirical' or 'falsifiable' if it divides the class of all possible basic statements unambiguously into the following two non-empty subclasses. First, the class of all those basic statements with which it is inconsistent [...]; and secondly, the class of those basic statements which it does not contradict [...].

Since Popper allows the use of auxiliary assumptions for the derivation of observation sentences (cf. Lakatos 1974, 106f), possible basic statements ('observation sentences' in the following) are those observation sentences compatible with the auxiliary assumptions. This leads to

Definition 1. A theory H is *falsifiable relative to* auxiliary assumptions A if and only if there is an observation sentence O with $A \nvDash \neg O$ such that $O \wedge A \vDash \neg H$.[4]

Here and in the following I will assume that for all theories H and auxiliary assumptions A it holds that $\Pr(H \wedge A) > 0$, and thus specifically that H and A are compatible ($H \wedge A \nvDash \bot$). Therefore, $H \wedge A$ must be compatible with some observation sentence, for otherwise it would entail all observation sentences and their negations. Thus definition 1 captures Popper's formulation.

The problem with falsifiability, Sober (2008, 130, cf. 151) points out, is that no purely probabilistic statement is falsifiable: "Consider a simple example: the statement that a coin has a probability of .5 of landing heads each time it is tossed [...] is testable, but it does not satisfy Popper's criterion". To avoid this result, Popper generalized his criterion, considering a theory H falsified even if an event occurs that is possible but very improbable according to $H \wedge A$. But this generalization runs into problems as well, the most important of which is, for the sequel, the following: If a theory implies an observation sentence, then it allows a deductive inference via *modus ponens*, '$\{X, X \rightarrow Y\} \vDash Y$': The assumption of the theory X and the fact that it implies Y entail Y. Popper justifies his criterion of falsifiability by using the implication of an observation sentence in a *modus tollens*, '$\{\neg Y, X \rightarrow Y\} \vDash \neg X$'. The justification of the probabilistic generalization of his criterion would thus have to rely on a probabilistic version of *modus tollens*, in which a theory X is false or at the very least improbable if $\Pr(Y \mid X)$ is high and Y is false. But as Sober (2002, 69f) points out (cf. Sober 2008, §1.4):

> There is a "smooth [t]ransition" between probabilistic and deductive *modus ponens*; the minor premiss ("X") either ensures that Y is true, or makes Y very

[4] I will follow Sober in treating observational claims and theories as single sentences, and the auxiliary assumptions as a finite set thereof. Finite sets of sentences are identified with the conjunctions of their members. Mostly (and always for the auxiliary assumptions), the restriction to single sentences—in effect a restriction to finite axiomatizations—is only a matter of notational convenience; I will note whenever it is essential. To allow sets of sentences and higher order logic in the definition of falsifiability, it must be phrased as "A theory H is falsifiable relative to assumptions A if and only if there is a set Ω of observation sentences with $\Omega \cup A \nvDash \bot$ such that $\Omega \cup H \cup A \vDash \bot$", where A and H are sets of sentences and $\bot$ is some contradiction.

> probable, depending on how the major premiss is formulated. In contrast, there is a radical discontinuity between probabilistic and deductive modus tollens. The *minor premiss* ("not-Y") guarantees that X is false in the one case, but has no implications whatever about the probability of X in the other.

Therefore, while Popper is right to infer from the fact that a theory entails an observation sentence that the theory is falsifiable, he cannot infer the same from the fact that it assigns a high probability to an observation sentence (and thus a low probability to the sentence's negation). Thus the generalization of falsifiability to probabilistic theories has not been justified.

A more successful criterion of empirical significance for probabilistic theories has been suggested within Bayesianism, the position that non-deductive inferences should follow the rules of the probability calculus, and specifically that the confirmation of scientific theories should follow Bayes's theorem,

$$\Pr(H \mid O \wedge A) = \frac{\Pr(O \mid H \wedge A) \cdot \Pr(H \mid A)}{\Pr(O \mid H \wedge A) \cdot \Pr(O \mid \neg H \wedge A)} \,. \tag{1}$$

$\Pr(H \mid A)$ is the probability of the theory given only the auxiliary assumptions, that is, before the observation O is taken into account, and thus called the prior probability (of H). $\Pr(H \mid O \wedge A)$ is the probability of H after O is taken into account, and hence called the posterior probability.[5] $\Pr(O \mid H \wedge A)$ and $\Pr(O \mid \neg H \wedge A)$ are the likelihoods of H and $\neg H$, respectively (for O). The standard criterion of empirical significance suggested within Bayesianism is the following (cf. Sober 2008, 150):

Definition 2. *Observations are relevant for* theory H *relative to* auxiliary assumptions A if and only if there is an observation sentence O such that

$$\Pr(H \mid O \wedge A) \neq \Pr(H \mid A) \,.^{6} \tag{2}$$

Sober (2008, 150, 24–30) argues that if H is, say, the theory of general relativity, it is well-nigh impossible to assign a probability to H, or assess the likelihood of $\neg H$, so that $\Pr(H \mid A)$, $\Pr(O \mid \neg H \wedge A)$, and $\Pr(H \mid O \wedge A)$ are often undefined. Bayes theorem (1) then becomes unusable, and definition 2 very questionable.

Sober (2008, 152) suggests a criterion of empirical significance that avoids all of the problems discussed so far. Unlike falsifiability, his criterion does not render all probabilistic

[5] Sober (2008, 8) calls $\Pr(H)$ the prior and $\Pr(H \mid O)$ the posterior probability and discusses Bayesianism without auxiliary assumptions. But he also argues that $\Pr(O \mid H)$, which would be used to determine $\Pr(H \mid O)$, is, unlike $\Pr(O \mid H \wedge A)$, almost never defined. Prior and posterior probabilities therefore have to be defined relative to auxiliary assumptions, lest Bayesianism be empty.

[6] I will always silently assume that for any occurring conditional probability $\Pr(B \mid C)$, $\Pr(C) \neq 0$.

theories empirically non-significant and does not rely on a faulty probabilistic generalization of *modus tollens* for its justification; unlike the Bayesian criterion, it does not rely on the probabilities of whole theories or on likelihoods of the negations of theories:[7]

> Hypothesis H_1 can now be tested against hypothesis H_2 if and only if there exist true auxiliary assumptions A and an observation statement O such that (i) $\Pr(O|H_1 \wedge A) \neq \Pr(O|H_2 \wedge A)$, (ii) we now are justified in believing A, and (iii) the justification we now have for believing A does not depend on believing that H_1 is true or that H_2 is true and also does not depend on believing that O is true (or that it is false).

Since 'now' is indexical, testability as defined is a predicate that changes over time. Sober (2008, 151) remarks that the "word 'now' marks the fact that whether a proposition has observational implications depends on the rest of what we are justified in believing, and that can change". This leads to a variety of inconveniences in connection with time operators. For instance, "Yesterday Marie declared: 'H_1 is testable against H_2'" can have a truth value different from "Yesterday Marie declared H_1 to be testable against H_2". The first sentence is true if and only if Marie was talking about the justificatory status of the auxiliary assumptions on the day of her utterance, the second sentence is true if and only if Marie was talking about the justificatory status of the auxiliary assumptions on the following day. It is therefore easier to think of contrastive testability as the three-place predicate 'H_1 can at time t be tested against H_2'. This still achieves Sober's intention, and arguably does so more explicitly.

The other indexical term, 'we', occurs only in the definiens, and thus makes Sober's criterion strictly speaking a claim rather than a definition. This is because the criterion violates the demand that in an explicit definition, any free variable of the definiens must also occur free in the definiendum, and thus Sober's criterion is creative (cf. Belnap 1993, 139): If for two theories H_1 and H_2, one referent of 'we' fulfills the definiens, the definiendum applies to H_1 and H_2. But then the definiendum applies to H_1 and H_2 no matter the referent, and thus any referent of 'we' fulfills the definiens for H_1 and H_2. This formal problem is easily solved by defining testability not only relative to time, but also relative to a group of people and thus as a four-place predicate.

As it stands, restriction (iii) on the auxiliary assumptions sounds like the demand that the justification of A must not depend on the fact that the truth of H_1, H_2, O or $\neg O$ is content of our belief. But very few statements are justified by the having of a belief, so that condition (iii) would be almost empty if this was meant. The restriction is therefore probably better expressed as the demand that the justifications for A must not depend on the fact that the belief in the truth of H_1, H_2, O or $\neg O$ *is justified*. For convenience, I

[7]Here and in the following quotations, I will replace Sober's '&' by '$\wedge$'.

will drop the reference to beliefs completely in the following, and speak only of justified sentences, rather than justified beliefs in the truth of propositions expressed by sentences.

Finally, the condition (i) on the likelihoods of H_1 and H_2 needs to be elucidated, given that Sober's critique of the Bayesian criterion of empirical significance assumes that some likelihoods are undefined. *Prima facie*, one would expect that H_1 and H_2 cannot be tested against each other if and only if $\Pr(O \mid H_1 \wedge A) = \Pr(O \mid H_2 \wedge A)$ for all O and A that fulfill conditions (ii) and (iii). But this would mean that the lack of testability is transitive for any theories that are not used to justify each other's auxiliary assumptions. And this is incompatible with Sober's remark that it is not clear that ID "can be tested against the Epicurean hypothesis that a mindless chance process gave vertebrates their eyes (or, for that matter, against the evolutionary hypothesis that the process of evolution by natural selection did the work)" (Sober 2008, 148). Assuming that the chance hypothesis can be tested against evolutionary theory (ET), if ID cannot be tested against either, the lack of testability is not transitive. The solution to this puzzle is that Sober (2010, 2f) interprets the inequality as true if and only if both likelihoods are defined and different. This interpretation, however, plays havoc with classical logic, for $p \neq q \mathrel{\dashv\vDash} \neg\, p = q$. Therefore, if the likelihoods p and q are defined and different, while the likelihood a is undefined, it follows from Sober's interpretation of the inequality that $p = a$ and $a = q$, while $p \neq q$. To avoid such inconsistencies, it is probably best to treat undefined likelihoods separately in the definition.

These considerations lead to

Definition 3 (Contrastive testability)**.** Theory H_1 can be *tested against* theory H_2 if and only if there are auxiliary assumptions A and an observation sentence O such that

(I) $\Pr(O \mid H_1 \wedge A)$ and $\Pr(O \mid H_2 \wedge A)$ are defined,

(II) $\Pr(O \mid H_1 \wedge A) \neq \Pr(O \mid H_2 \wedge A)$,

(III) A is justified, and

(IV) the justification of A

   a) does not depend on H_1 or H_2 being justified and
   b) does not depend on O or $\neg O$ being justified.

One could reformulate definition 3 to include a reference to time and groups of people, that is, define 'H_1 can be tested against H_2 *at time t by group g*' by relativizing 'justification' (and possibly 'dependence') to time t and group g. In similar cases, especially when the auxiliary assumptions are simply the background assumptions, these relativizations are typically suppressed as it is clear that the background assumptions and generally the set of justified sentences can change over time and from group to group. Thus I will do likewise.

Sober calls his criterion simply 'testability', but the qualifier 'contrastive' distinguishes it clearly from the ordinary language term and emphasizes that, atypically, the empirical significance of one theory is defined relative to another. It may seem problematic to explicate a one-place predicate like 'makes observational assertions' by a two-place predicate like contrastive testability. Frege (1918, 291), for example, objects to the explication of 'truth' as a correspondence relation on the grounds that the first is a one-place, the second a two-place predicate. However, many successful explications involve a change of the logical structure, as the explication of 'warm' by 'warmer than' and finally 'temperature' illustrates (Carnap 1950, §4). Hempel (1952, §10) even argues that the move from a classificatory to a comparative concept is often a sign of an investigation's maturity and, at another point, argues that empirical significance is a matter of degree (Hempel 1965, 117):

> Significant systems range from those whose entire extralogical vocabulary consists of observation terms, through theories whose formulation relies heavily on theoretical constructs, on to systems with hardly any bearing on potential empirical findings.

It is plausible that such an explication of a classificatory by a comparative concept is helpful (Popper 1935, §33; Lutz 2010a, §8.1). But unlike 'warmer than' and 'more significant than', contrastive testability is symmetric: The definiens is invariant up to logical equivalence under exchange of H_1 and H_2. This spells trouble for the intended use of contrastive testability, the debate between proponents of ET and ID. Specifically, it means that if "the hypothesis of intelligent design cannot be tested against evolutionary theory, at least at present" (Sober 1999, 64), then ET also cannot be tested against ID. Thus the lack of contrastive testability is, ceteris paribus, bad for ID if and only if it is bad for ET. The ceteris paribus condition is necessary because there may be other criteria that break the symmetry; contrastive testability itself, however, provides no reason to prefer ET over ID. Specifically, if the lack of contrastive testability expresses that ID "doesn't predict much of anything" (Sober 2008, §2.14), then the same holds for ET.

In some passages Sober himself uses 'testability' like a one-place predicate. For instance, his claims that 'Undetectable angels exist' is untestable and that 'This coin has probability of .5 of landing heads' is testable are, strictly speaking, meaningless according to his own definition. And both claims are important for Sober's line of argument, since he relies on the first to argue that testability is different from meaningfulness, and on the second to argue that falsifiability is not an adequate criterion of empirical significance.

3 The restrictions on the auxiliary assumptions

Before looking further at the symmetry of definition 3, it is helpful to investigate some of its other features. Justus (2010) already notes the imprecision of the terms 'justified', 'belief', and 'depend', and the problems introduced by the time-index 'now'. I will try to stay clear of the pitfalls that come with these terms when discussing additional problems with the restrictions (IVa) and (IVb) on the auxiliary assumptions.

Sober (1999, 54) introduces auxiliary assumptions into the definition because "hypotheses rarely make observational predictions on their own; they require supplementation by auxiliary assumptions if they are to be tested" (cf. Sober 2007, 5f; Sober 2008, 144).[8] But this "raises the question of which auxiliary assumptions we should use to render a theory testable. What makes an auxiliary assumption 'suitable'?" (Sober 2008, 144). He justifies the restrictions (IVa) and (IVb) as follows (Sober 2008, 145; cf. Sober 2007, 6):

> I hope it is obvious that if you want to use the observation O to test hypothesis H_1 against hypothesis H_2, that the auxiliary assumptions you make must not depend for their justification on assuming that H_1 is true or on assuming that H_2 is true. What is perhaps less obvious is that the auxiliary assumptions must be justified without assuming that O is true. Here is why that additional constraint is needed: If O is true, so is the disjunction "either H_1 is false or O is true". If you use this disjunction as your auxiliary assumption A_1, then it turns out that the conjunction $H_1 \wedge A_1$ entails O. This allows H_1 to make a prediction about O even when H_1 has nothing at all to do with O. The same ploy can be used to obtain auxiliary assumptions A_2 so that the conjunction $H_2 \wedge A_2$ also entails O. Using propositions A_1 and A_2 as auxiliary assumptions leads to the conclusion that the two hypotheses H_1 and H_2 both have likelihoods of unity.

As it stands, this argument proves nothing about the relevance of restriction (IVb) for the definition of contrastive testability, since it only shows that for one specific auxiliary assumption, $A \boxminus A_1 \wedge A_2$, both theories' likelihoods are 1. But to show that H_1 cannot be tested against H_2, their likelihoods have to be identical for *all* auxiliary assumptions that fulfill restrictions (III) and (IVa). Furthermore, if the goal was to arrive at the *same* likelihood for both theories, $A \boxminus O$ would achieve the same result.

But the ingenuity of the choice of A_1 is exactly that, if H_1 and H_2 have nothing at all to do with O, the likelihood of $H_1 \wedge A_1$ is 1, while the likelihood of $H_2 \wedge A_1$ is not. Reconceptualized in this way, Sober's case for restriction (IVb) is a typical trivialization proof, since it shows that without it, any two theories can be tested against each other. The argument has three implicit assumptions. Sober's first assumption about justification is that a sentence S (here: $\neg H_1 \vee O$) logically entailed by a justified sentence J (here: O) is also justified, since otherwise S might be excluded by restriction (III). Sober's second assumption about justification is that S depends for its justification only on J, for otherwise, S might be excluded by restriction (IVa) or, implausibly, for its dependence on $\neg O$ by

[8] Since Sober does not use 'prediction' to refer exclusively to claims about the future, I will treat it as synonymous with 'assertion'.

restriction (IVb). The third assumption of the proof is that H_1 is not a tautology, since otherwise $A_1 \boxminus O$.

Sober does not show why the reference to $\neg O$ in restriction (IVb) is necessary, but it is clear that otherwise Sober's trivialization proof could be repeated: Assume, as Sober does,[9] that the negation of an observation sentence is itself an observation sentence. Since $\Pr(O \,|\, H_i \wedge A) = 1 - \Pr(O \,|\, H_i \wedge A), i \in \{1, 2\}$, A can then be justified with O, while the likelihoods of H_1 and H_2 would differ for $\neg O$.

Sober claims that restriction (IVa) is obvious, and that without it, the criterion would beg the question (Sober 2007, 6). But this is at least not obvious. Arguments must not in general allow their conclusion among their premises (that is, beg the question) because otherwise every claim could be shown to be true, and the concept of an argument would be trivial. But even without restriction (IVa), it is not possible to simply assume that H_1 can be tested against H_2 when the criterion is applied. In fact, I want to show that (IVa) is often redundant, and in the remaining cases it is either ineffective or lacks a justification.

As for the redundancy, note that restriction (III) demands that A has to be justified, and that a sentence whose justification depends on another sentence B is justified only if B is justified. Giving up this relation between 'justified' and 'depend' would render Sober's restriction (IV) altogether empty. Thus an auxiliary assumption whose justification depends on H_1 can only be used in definition 3 if H_1 itself is justified, and analogously for H_2. Typically, however, the question of empirical significance does not even come up for theories that are already justified. Indeed, Sober assumes that a theory is confirmed only if it is tested, and this is possible only if it is contrastively testable. Assuming that only confirmed theories are justified, restriction (IVa) therefore goes beyond restriction (III) *only* when the question of empirical significance has already been answered.[10]

Restriction (IVa) is ineffective at least in the following cases: Assume that for some O and A, $\Pr(O \,|\, H_1 \wedge A) \neq \Pr(O \,|\, H_2 \wedge A)$, and the justification of A depends on H_2. Depending on the exact explication of 'justification', it will often be the case that A depends for its justification not on H_2 in its full logical strength, but can rather also be justified by a logically slightly weaker theory H'_2, which may, for example, result from H_2 by a slight decrease of the domain of applicability. Or, if A does depend on H_2 in it full logical strength, there may be slightly weaker auxiliary assumptions A' that lead to the same likelihoods but depend only on H'_2. Since it is implausible that H'_2 depends on H_2 for its justification (it is rather the other way around), H_1 and H_2 are contrastively testable in spite of restriction (IVa).

In general it is not obvious why, if they are justified, H_1, H_2, or the sentences that depend on them for justification should be excluded from the auxiliary assumptions. The

[9]See the discussion on page 20.

[10]Of course, one may want to justify or confirm an already justified or confirmed theory *further*, but this is then not a question of contrastive testability any more.

example that Sober (2008, 145) adduces to show the need for restriction (IVa) does not make his case, as I will argue:

> The point [. . .] is that, in testing H_1 against H_2, you must have a reason to think that the auxiliary proposition A is true that is independent of whatever you may already believe about H_1 and H_2. For example, suppose you are on a jury. Jones is being tried for murder, but you are considering the possibility that Smith may have done the deed instead. Evidence is brought to bear: A size 12 shoe print was found in the mud outside the house where the murder was committed, as was cigar ash, and shells from a Colt .45 revolver. Do these pieces of evidence favor the hypothesis that Smith is the murderer or the hypothesis that Jones is? It is a big mistake to answer these questions by *inventing* assumptions. If you assume that Smith wears a size 12 shoe, smokes cigars, and owns a Colt .45 and that Jones wears a size 10 shoe, does not smoke, and does not own a gun, you can conclude that the evidence favors Smith over Jones.

First note that in this example the question is which theory can be inferred from the evidence, not which observations are asserted by the theory; that is, the example revolves around a question of confirmation, not empirical significance. More important in the following is that the belief about Smith's shoe size would be excluded from the auxiliary assumptions even without restriction (IVa). This is because, first, the belief that Smith is the murderer (H_1) is itself not justified, and thus cannot justify anything. Second, Smith's murdering someone does not allow to conclude anything about her shoe size. The conclusion also requires the belief that there was a size 12 shoe print at the crime scene (O). In other words, the justification of the auxiliary assumption depends on H_1, so that it is excluded from A by restriction (III), and depends on O, so that it is excluded by restriction (IVb).

Restriction (IVa) is included in definition 3 for more serious reasons than fictitious murder trials with careless jurors. It is meant to address an argument in defense of the testability of ID that Sober (1999, 65, note and line break removed) describes as follows (cf. Sober 2008, 143–146):

> [A]dvocates of the design argument should not be confident that they know what characteristics God would have wanted to give to organisms on earth if he had created them. Creationists may be tempted to respond to this challenge simply by inspecting the life we see around us and saying that God wanted to create *that*. After all, if life is the result of God's blueprint, can't we infer what the blueprint said by seeing what the resulting edifice looks like? [But you] can't just *assume* that God created organisms, and you also can't *assume* that if God created organisms he would have made them with such-and-such characteristics.

Analogous to the murder trial, the justification of the auxiliary assumption about God's characteristics in the creationists' argument depends both on the assumption that God exists and the observational assumption that life is as we see it around us, like "that".[11] Therefore it is excluded from A by restriction (III) because it is not justified until belief in God is justified. And if a description of life as we see it around us is given as the observation sentence O for which the likelihoods of ID and ET differ, then the auxiliary assumption is also excluded by restriction (IVb) for dependence on O.

Let me now turn to Sober's justification of restriction (IVb) and show that it is not sufficient to avoid trivialization of definition 3. Specifically, any two theories can be tested against each other as long as one of them can be finitely axiomatized and thus phrased as one (possibly inordinately long) sentence:[12]

Claim 1. *Let H_1 and H_2 be theories and O and S be any two sentences such that*

1. *O is an observation sentence,*
2. *$S \vDash O$,*
3. *S is justified independently of O, $\neg O$, H_1, or H_2,*
4. *$\Pr(O \mid H_1 \wedge A)$ and $\Pr(O \mid H_2 \wedge A)$ are defined, and*
5. *$\Pr(O \mid H_2 \wedge \neg H_1) \neq 1$.*

Under Sober's assumptions about justification, H_1 and H_2 can then be tested against each other.

Proof. Choose O and S such that conditions 1–5 hold. Then condition (I) of definition 3 is fulfilled. Since S is justified, so is $A \,\vDash\!\!\dashv\, \neg H_1 \vee S$ by Sober's first assumption about justification. It follows from Sober's second assumption about justification that, since the justification of S does not depend on O, $\neg O$, H_1, or H_2, neither does the justification of A. Therefore A fulfills restrictions (III) and (IV) of definition 3. Since $\Pr(O \mid H_1 \wedge A) = 1$ and from $\Pr\big(O \mid H_2 \wedge \neg H_1\big) \neq 1$ and $S \vDash O$ it follows that $\Pr\big(O \mid H_2 \wedge (\neg H_1 \vee S)\big) \neq 1$ (see appendix, claim 10), it holds that $\Pr(O \mid H_1 \wedge A) \neq \Pr(O \mid H_2 \wedge A)$. H_1 and H_2 therefore fulfill condition (II) of definition 3 and can be tested against each other. □

Note that for $\Pr(H_1 \wedge H_2) = 0$, condition (5) simplifies to '$\Pr(O \mid H_2) \neq 1$' and that the trivialization proof that Sober uses to justify restriction (IVb) can be recovered by dropping the independence from O in condition (3) and choosing $S = O$.

[11] In disanalogy to the murder trial, the question in this case is indeed which observations the theory asserts, not what can be inferred from the observations.

[12] Since the proof relies on the use of the negation of one of the theories, this restriction is essential.

Conditions 1–3 are impossible to fulfill if a justification can proceed only deductively from observation sentences, because then the justification of a sentence depends on every observation sentence it entails. However, since Sober's criterion is meant to be applicable to inductive theories, it is plausible that auxiliary assumptions can also be inductively justified. In that case, it is easy to find sentences S and O that fulfill all the requirements. For instance, let S express that a specific vase does not break when dropped a hundred times from a specific height, and O express that the vase does not break on the hundredth drop. Then S is justified independently of O when the vase is dropped 99 times without breaking, so that S and O fulfill conditions 1–5 for any two theories that are not about vases. Since according to Sober (1999, 54), "hypotheses rarely make observational predictions on their own", that includes almost all theories. But even for two theories that make assertions about vases, it should not be difficult to find other observations that neither they nor their negations assert with probability 1, but that can be asserted by enumerative induction.

To summarize, Sober's restriction (IVa) is unjustified where it is not redundant or ineffective, and restrictions (III) and (IV) together are to weak to avoid trivialization. Clearly, the search for general restrictions on the auxiliary assumptions poses a host of subtle problems. To bracket these problems, I suggest

Definition 4. Theory H_1 can be *tested against* theory H_2 *relative to* auxiliary assumptions A if and only if there exists an observation sentence O such that $\Pr(O \mid H_1 \wedge A)$ and $\Pr(O \mid H_2 \wedge A)$ are defined and

$$\Pr(O \mid H_1 \wedge A) \neq \Pr(O \mid H_2 \wedge A)\ . \tag{3}$$

This definition makes it necessary to decide on a case by case basis which auxiliary assumptions are suitable. This may be a good preliminary strategy, because often the suitable auxiliary assumptions are, under the moniker 'background assumptions', reasonably clear.

Eventually, of course, it would be helpful to have a general criterion for suitable auxiliary assumptions and define absolute contrastive testability as contrastive testability relative to suitable auxiliary assumptions. To this end, I suggest the following. The proof of claim 1 is a modification of a trivialization proof for Sober's criterion of 'having observable implications' (Lutz 2010a, §9.2) and leads to a similar diagnosis. Sober's proof and that of claim 1 rely on the possibility to include a sentence ($\neg H_1 \vee O$ or $\neg H_1 \vee S$) in A that is justified by another one (O or S) that is itself not included in A. Both trivialization proofs can therefore be blocked by explicating 'suitable auxiliary assumptions' as 'honest set of auxiliary assumptions':

Definition 5. A is an *honest set* if and only if every $S \in A$ is a justified sentence, and A also contains every sentence on which the justification of S depends.

Note that this definition only uses concepts that already occur in Sober's definition 3 of contrastive testability and that A can be a proper subset of the set of all justified sentences.

Definition 5 allows to modify Sober's criterion of testability as follows:

Definition 6. Theory H_1 can be *tested against* theory H_2 if and only if H_1 can be tested against H_2 relative to an honest set of auxiliary assumptions.

To distinguish clearly between concepts that are defined relative to auxiliary assumptions (as in definition 4) and those that are not (as in definition 6), I will refer to the latter sometimes as *absolute* concepts.

The restriction of the auxiliary assumptions in definition 6 to honest sets entails restriction (III) of definition 3. And while the restriction to honest sets does not entail restriction (IVb), it precludes all trivializations precluded by that restriction: Two theories H_1 and H_2 fail to be contrastively testable because of restriction (IVb) only if for any S whose inclusion in A would lead to differing likelihoods for some O, the justification of S depends on O or $\neg O$. In that case, (IVb) ensures that H_1 and H_2 are not contrastively testable. The restriction of A to honest sets leads to the same result, because if the justification of S depends on O ($\neg O$) and A is honest, then $O \in A$ ($\neg O \in A$). Thus $\mathrm{P}(O \mid H_1 \wedge A) = 1 = \mathrm{P}(O \mid H_2 \wedge A)$ ($\mathrm{P}(O \mid H_1 \wedge A) = 0 = \mathrm{P}(O \mid H_2 \wedge A)$). As an example, take the sentence $\neg H_1 \vee O$ of Sober's trivialization proof. Restriction (IVb) excludes $\neg H_1 \vee O$ from A, so that the likelihoods of H_1 and H_2 for O cannot differ because of $\neg H_1 \vee O$. The restriction to honest sets, on the other hand, leads to the inclusion of O in A, so that the likelihoods do not differ, either. Unlike restriction (IVb), the restriction to honest sets also leads to identical likelihoods if $\neg H_1 \vee S \in A$ is justified by a sentence $S \vDash O$, thereby precluding the proof of claim 1.

Since it is not clear in which case restriction (IVa) is meant to preclude trivialization, or in general, which problem it is meant to solve, I cannot show that definition 5 can fulfill the role of restriction (IVa). Given the restriction's questionable role and justification, this should not be considered a drawback of definition 6. If there is a justification for restriction (IVa), however, one can modify definition 6 by defining contrastive testability as contrastive testability relative to an honest set that does not include H_1 or H_2. This restriction entails restriction IVa.

While definition 6 avoids the above two trivialization proofs, it might allow others. In response to such a proof, one can fall back on definition 4 until a better explication of 'suitable' is found than definition 5. In general, any results on which auxiliary assumptions are suitable, or which assumptions are background assumptions, can be used directly as a substitute for definition 5 (cf. Lutz 2010a, §9.2).

4 Conditions of adequacy for a criterion of empirical significance

Sober's definition of contrastive testability, definition 3, is inadequate as a criterion of empirical significance because it is trivial. Of course, there are other conditions of adequacy that a criterion of empirical significance has to fulfill. I will argue that Sober's assumptions and his intended application of the criterion lead to eight such conditions.

Many of these conditions of adequacy relate empirical significance to concepts that rely on assertion, that is, a kind of inference, and therefore have a deductive and a probabilistic formulation. This is because deductive inference (entailment: $B \vDash C$) clearly does not generalize probabilistic inference (e. g., $\Pr(B) = q$ and $\Pr(C \mid B) = 1$, thus $\Pr(C) = q$). But probabilistic inference also does not generalize deductive inference. For assume that the domain has infinite cardinality. Then it may be that $\Pr(C \mid B) = 1$, but there are cases in which B is true and C is false. This happens, for example, when the domain is the interval $[0, 2]$ with a uniform probability distribution, B is '$x \leq 1$', and C is '$x < 1$' (cf. Feller 1971, 33f).

This difference between the deductive and the probabilistic concept of inference generally leads to differences between the deductive and probabilistic formulations of the conditions of adequacy, which in turn may lead to one criterion of deductive empirical significance (in the following sometimes shortened to 'deductive criterion') and a separate criterion of probabilistic empirical significance ('probabilistic criterion'). A theory may then be called empirically significant (simpliciter) if and only if it is deductively or probabilistically empirically significant.

4.1 Conditions of adequacy for a criterion of empirical significance

(A) The criterion should not be trivial. A trivial definition, one that includes all or no objects of the domain, cannot be a good explicatum for a concept that is meant to include some, but not all objects of the domain. At the very least, a trivial explicatum is uninformative. Since Sober intends to distinguish between theories that are worthy to be pursued and theories that are not, his criterion must not be trivial, and he implicitly relies on this condition of adequacy when arguing for restriction (IVb) of definition 3. Like the proof of claim 1, his argument assumes that the theories under scrutiny are not tautologies. A criterion of empirical significance that excludes or includes only tautologies is trivial only in the domain of contingent theories, but this is arguably trivial enough for such a criterion to be inadequate.

(B) The criterion should include all and only theories that make observational assertions. Popper (1935) justifies his criterion of falsifiability with the assumption that all and only theories that make observational assertions are empirically significant. The assumption is justified by Ayer (1936, 97), who argues that "the purpose" of an empirical theory is "to enable us to anticipate the course of our sensations". If Ayer's argument is sound, empirical significance is a necessary and sufficient condition for making observational assertions.

Sober (2008, 130) states that "a testable statement makes predictions, either by deductively entailing that an observation will occur or by conferring a probability on an observational outcome." Thus for Sober empirical significance is a sufficient condition for the making of observational assertions. Let this be condition (i). Sober also subscribes to the converse of condition (i) as can be seen from his claims that "[t]he problem with the hypothesis of intelligent design is [...] that it doesn't predict much of anything" (Sober 2008, §2.15) and that his "criticism of the design argument might be summarized by saying that the design hypothesis is untestable" (Sober 2008, 148). However, since Sober (2008, §2.12) infers the lack of empirical significance from the lack of observational assertions, his criticism of ID *relies* only on condition (i). Sober's criticism of Popper's falsifiability criterion does seem to rest on the converse of condition (i): 'This coin has probability of .5 of landing heads each time it is tossed' makes a probabilistic assertion, and its lack of falsifiability is a reason for Sober to reject Popper's criterion. This seems to assume that every theory that makes probabilistic assertions is empirically significant. Therefore, Sober's arguments against ID and falsifiability rest on the condition that all and only theories that make observational assertions are empirically significant.

Sober (2008, 52, n. 29) further states two relations between deductive empirical significance and the making of deductive observational assertions:

> If a true observation sentence entails H [...] you can conclude without further ado that H is true; this is just *modus ponens*. And if H entails O and O turns out to be false, you can conclude that H is false [...]; this is just *modus tollens*.

These are two sufficient conditions for empirical significance, (ii) entailment *by* an observation sentence and (iii) entailment *of* an observation sentence. Condition (iii) is the converse of condition (i) when inferences are restricted to entailment (cf. Sober 1999, 72, n. 14). Thus, as argued above for inferences in general, a theory makes deductive observational assertions if and only if it is deductively empirically significant.

Condition (ii), however, is arguably incompatible with condition (i): For any sentence S and observation sentence O, $O \vDash O \vee S$, that is, according to condition (ii), $O \vee S$ is empirically significant. But let S be such that it does not make observational assertions, that is, for any observation sentence O', $S \nvDash O'$, and S does not confer any probability on O'. Then, as a matter of logic, $O \vee S \nvDash O'$, so $O \vee S$ does not make deductive assertions. Arguably, $O \vee S$ also does not confer a probability on any observation sentence. If, for example, $O \vee S$ confers probability x on O' if and only if $O \vee S \vDash \Pr(O') = x$, then the argument for the deductive case can be repeated. If O and S are chosen so that they themselves are not assigned probabilities, then it is plausible that for any O', $\Pr(O' \mid O \vee S)$ also cannot be assigned a probability. Thus $O \vee S$ does not make any observational assertions and is thus not empirically significant according to condition (i), which is incompatible with condition (ii). On pain of inconsistency, Sober therefore has to choose whether all theories entailed by observation sentences are empirically significant or whether all theories that are empirically significant make observational assertions. Given that his core argument against ID is that ID fails to make assertions, I take it that he would choose the latter.[13]

(C) The criterion should exclude all theories that are empirically equivalent to tautologies. What it means for a theory to make probabilistic assertions may not be completely clear, especially since some probabilities may be undefined. It will therefore be convenient to also have a plausible corollary of condition (B), starting from the observation that a tautology $\top$ makes no deductive assertions, since $B \wedge \top \vDash C$ only if $B \vDash C$, and makes no probabilistic assertions either, since $\Pr(C \mid B \wedge \top) = \Pr(C \mid B)$ for all B and C. Therefore, tautologies should be excluded by any criterion of empirically significance.

Flew (1950, 258) goes so far to call every theory that does not make observational assertions a tautology, but this is clearly to strong: 'Borogroves are mimsy' is not a tautology, but on account of containing two undefined terms, does not make any observational assertions. Rather, any theory that makes the same observational assertions as a tautology should be taken as not empirical significant. Unlike condition of adequacy (B), which relies on some criterion for the making of observational assertions, condition (C) relies on a criterion for empirical equivalence, the making of *the same* observational assertions. Out of caution, one may treat the empirical non-equivalence to a tautology as a necessary, but not as a sufficient condition for empirical significance.

(D) The criterion should allow some auxiliary assumptions. In logic and mathematics, the power of using justified auxiliary assumptions is well known. A previously established theorem can always be used in a proof, and by doing so, the proof is often enormously shortened. In the sciences, justified auxiliary assumptions are often other theories or boundary conditions for a specific experiment, and disallowing either would make scientific research impossible. Sober (1999, 54) follows Duhem (1914) in stressing the latter point when he states that "hypotheses rarely make observational predictions on their own" (cf. Sober 2007, 5f; Sober 2008, 144). Therefore, theories that rely on suitable auxiliary assumptions should not thereby be automatically empirically non-significant.

[13]Note that the claim "There is an intelligent designer" is equivalent to "There is a human designer or there is a non-human designer" and thus analytically entailed by an observation sentence like "There are humans who design". Arguably, however, "There is a non-human designer" does not make an observational assertion, so that "There is an intelligent designer" does not either. For further discussion, see Lutz (2010b, §4.1).

(E) The criterion should not exclude all probabilistic theories. A probabilistic sentence is, in a straightforward way, less certain than a non-probabilistic one. But that does not mean that a probabilistic sentence is therefore of no use. In fact, decision theory has been developed on the basis of probabilistic sentences about outcomes of actions, and its practical uses are obvious. Thus a theory that makes probabilistic assertions should not be rendered empirically non-significant. For Sober (1999, 57), this condition of adequacy follows from condition (B) and the fact that probabilistic assertions are also assertions (cf. Sober 2007, 5; Sober 2008, 130).

(F) The criterion should not rely on the probabilities of whole theories or likelihoods of the negations of whole theories. Sober (2008, 24–30) argues that for many theories H the probabilities $\Pr(H \mid A)$, $\Pr(H \mid O \wedge A)$, and $\Pr(O \mid \neg H \wedge A)$ are undefined (cf. Sober 1990, §III). A criterion that relies on these probabilities would therefore be unusable in many cases.

(G) The criterion should be equivalent to a Bayesian criterion of empirical significance whenever all occurring probabilities are defined. Since Bayesianism relies on probabilities of whole theories and likelihoods of negations of whole theories, Sober rejects it as a general method of scientific inference. Instead, Sober (2008, 37) suggests *likelihoodism*, which relies only on the likelihoods of a theory, but notes (cf. Sober 2008, 32):

> The likelihoodist is happy to assign probabilities to hypotheses when the assignment of values to priors and likelihoods can be justified by appeal to empirical information. Likelihoodism emerges as a statistical philosophy distinct from Bayesianism only when this is not possible.

Since there are criteria of empirical significance that have been developed within Bayesianism, this suggests that a probabilistic criterion of empirical significance should be equivalent to one of these Bayesian criteria whenever all probabilities are defined. This Bayesian criterion should, of course, fulfill all other criteria of adequacy.

(H) The probabilistic criterion should contain as a special case an adequate criterion of deductive empirical significance that relies only on *modus ponens*. As already mentioned, Sober sees a smooth transition between probabilistic and deductive *modus ponens* (Sober 2002, 69f). More specifically, Sober (2008, 50) points out that the following holds:

(Update) $\Pr_{\text{then}}(H \mid O)$ is very high
O
O is all the evidence we have gathered between then and now.
—————————————————————
$\Pr_{\text{now}}(H)$ is very high

> This is nothing other than the rule of updating by strict conditionalization. (Update) is a sensible rule, and it also has the property of being a generalization of deductive modus ponens.

As argued at the beginning of this section, (Update) is not, strictly speaking, a generalization of *modus ponens*. But at least when all and only sentences with probability 1 are certain, deductive and probabilistic inference coincide. This can be put more precisely as follows. Each structure $\mathfrak{M}$ of a language $\mathscr{L}$ of predicate logic assigns a truth value to each sentence in $\mathscr{L}$. If $\Pr_{\mathfrak{M}}$ is defined as the function that assigns 1 to all sentences true in $\mathfrak{M}$ and 0 to all sentence false in $\mathfrak{M}$, then $\Pr_{\mathfrak{M}}$ is a probability assignment (see appendix, claim 11). Call such probability assignments *truth value-like*. For truth value-like probability assignments, probabilistic inferences and deductive inferences coincide: The possible values of $\Pr(C \mid B)$ are restricted to 0 and 1, and $\Pr(C \mid B) = 1$ if and only if $B \vDash C$ (as always assuming that $\Pr(B) \neq 0$; see appendix, claim 12). Truth value-like probabilities may be assigned by fiat, but they also occur more or less naturally when the domain of the theory is finite (so that there cannot be claims with probability 1 that are not certain) and there are no regularities whatsoever, so that no probabilities can be assigned to statements about the domain that are not known to be true.[14]

In this sense, then, there can be a smooth transition between probabilistic and deductive inference. Given that all and only theories that make deductive or probabilistic assertions must be empirically significant by condition of adequacy (B), there must then also be a smooth transition between any criterion of probabilistic empirical significance and a criterion of deductive empirical significance that uses the implications of the theory only in a *modus ponens*. As I will say, the probabilistic criterion must contain as a special case a deductive criterion that relies only on *modus ponens*. Of course, the deductive criterion should fulfill all those conditions of adequacy that also have purely deductive formulations, that is, conditions (A), (B), (C), and (D). To fulfill condition (B), it is enough for the deductive criterion to include all and only theories that make deductive assertions, because it is impossible that it could include theories that make only probabilistic assertions. Analogously, it is enough if the criterion excludes all theories that are deductively empirically equivalent to a tautology to meet condition (C).

Independently of any smooth transition in the case of assertions, the criterion of empirical significance simpliciter should be a generalization of an adequate deductive criterion. Thus, when the deductive and the probabilistic criterion coincide, the probabilistic criterion must not include theories that the deductive criterion excludes. For if it did, these theories would be included by the criterion of empirical significance simpliciter, and thus this criterion would not generalize the deductive criterion, but rather contradict it.

* * *

[14] The latter is arguably the case in Popper's approach to induction (cf. Salmon 1967, §II.3).

Some of these conditions of adequacy are controversial; especially condition (F) would be challenged by Bayesians. But these conditions all follow from Sober's basic assumptions or apply to Sober's criterion because of its intended application. Of course, it may be that these conditions of adequacy are incompatible, so that some have to be given up. This is the case for conditions (i) and (ii) discussed under condition of adequacy (B). But a criterion of empirical significance that is to be applied as Sober intends should fulfill as many of these conditions as possible.

4.2 Falsifiability is the unique adequate criterion of deductive empirical significance

Condition of adequacy (H) demands that the deductive criterion contained in the probabilistic criterion can be phrased in terms of *modus ponens* and fulfills all those conditions of adequacy that pertain to deductive criteria. In this section, I want to show that these conditions uniquely determine falsifiability. According to definition 1, a theory H is falsifiable relative to auxiliary assumptions A if and only if there is an observation sentence O with $A \nvDash \neg O$ such that $O \wedge A \vDash \neg H$. This can be combined with definition 5:

Definition 7. A theory H is *falsifiable* if and only if it is falsifiable relative to an honest set.

Sober (1999, 48–57) defends many of the assumptions on which falsifiability depends against criticisms (see also Lutz 2010a, §2). As for the conditions of adequacy:

To show that a criterion is not trivial and thus fulfills condition (A), it is enough to give a positive and a negative instance of the criterion. If O is an observation sentence, then O is falsifiable relative to $\varnothing$, and if S is a sentence none of whose terms are observational, then S is not falsifiable relative to $\varnothing$. Thus definition 1 is not trivial. If O is an observation sentence, then O is also absolutely falsifiable according to definition 7, for it is falsifiable relative to $\varnothing$, which is an honest set according to definition 5. Thus there is a falsifiable sentence. I will not attempt to prove that there is a non-falsifiable sentence, because this would amount to finding a sentence that is not falsifiable relative to any honest set. The proof is immediate for tautologies, but impossible for contingent theories without more precise notions of justification and dependence.

That falsifiability includes all theories that make deductive assertions is already pointed out by Sober (1999, n. 14), who remarks that if "$H \wedge A$ deductively entails O, and A is known to be true, then, if we observe not-O, we can conclude that H is false." Note that Sober here assumes, like Popper, that the negation of an observation sentence is again an observation sentence. Under this assumption, condition (B) uniquely determines the criterion of relative falsifiability, if it is further assumed that a theory makes deductive assertions if and only if it entails observation sentences not entailed by the auxiliary assumptions alone:

Definition 8. A theory H *makes deductive observational assertions relative to* assumptions A if and only if there is an observation sentence O such that $H \wedge A \vDash O$ and $A \nvDash O$.[15]

Definition 8 is equivalent to Sober's definition of 'having observational implications' (Sober 2008, 151) except that the problem of suitable auxiliary assumptions is bracketed. Sober uses restrictions (III) and (IV) to determine the auxiliary assumptions, which, as mentioned, allows a trivialization proof.

Definition 8 leads to

Claim 2. *If the negation of an observation sentence is again an observation sentence, then a theory H is falsifiable relative to A if and only if it makes deductive observational assertions relative to A.*

Proof. If O, A, and H are sentences, the proof is immediate. If O, A, and H are sets, the claim follows immediately from claim 13 (see appendix). □

Note that the condition on observation sentences is not only implicitly assumed by Sober, but also fulfilled by the most common restrictions on observation sentences: The negation of an observation sentence is itself observational if all and only sentences with a specific non-logical vocabulary are observational (cf. Psillos 2000, 158f), if all and only molecular sentences with a specific vocabulary are observational (cf. Carnap 1937, §23), and if all and only sentences are observational whose quantifiers are relativized to observable objects (cf. Carnap 1956, §II; Friedman 1982, 276f). A sentence could also be considered observational if and only if it is about observations, and according to Lewis (1988, 140f), if a sentence is about observation, so is its negation. All these restrictions even entail that the set of observation sentences is closed under truth-functional composition. Therefore relative falsifiability arguably meets condition of adequacy (B).

Claim 2 establishes that absolute falsifiability meets condition (B) if and only if it is the case that a theory makes deductive assertions iff it makes deductive assertions relative to an honest set. But even if the definition of an honest set turns out to be wanting in some respect, there is no obvious reason to doubt that the auxiliary assumptions suitable for falsifiability are also suitable for assertions. Rather, since background assumptions are usually considered to be independent from the concepts that rely on them, this is a fairly plausible conjecture. Under this conjecture, all equivalence results between relative concepts transfer to absolute concepts, and it will be silently assumed in the following.

[15] To allow sets of sentences and higher order logic, the definition must be phrased as "A theory H makes deductive observational assertions relative to assumptions A if and only if there are a set Ω of observation sentences and an observation sentence O such that $\Omega \cup H \cup A \vDash O$ and $\Omega \cup A \nvDash O$". If a logic is compact, $\Omega \cup H \cup A \vDash O$ if and only if there is a finite set Ω' such that $\Omega' \cup H \cup A \vDash O$, which is equivalent to $H \cup A \vDash \neg(\bigwedge \Omega' \wedge \neg O)$. Hence this definition reduces to definition 8 in first order logic if the set of observation sentences is closed under truth-functional composition.

Independently of the question of auxiliary assumption, claim 2 establishes that both falsifiability and relative falsifiability can be phrased so that they rely only on *modus ponens*.

Condition of adequacy (C) is fulfilled because no theories that are deductively empirically equivalent to a tautology $\top$ are falsifiable:

Definition 9. Two theories H_1 and H_2 are *deductively empirically equivalent relative to* assumptions A if and only if for all observation sentence O, $H_1 \wedge A \vDash O$ iff $H_2 \wedge A \vDash O$.[16]

Claim 3. *If the negation of an observation sentence is again an observation sentence, H is falsifiable relative to A if and only if H and $\top$ are not deductively empirically equivalent relative to A.*

Proof. H is not deductively empirically equivalent to $\top$ if and only if there is an observation sentence O such that either $H \wedge A \vDash O$ and $A \nvDash O$ or $H \wedge A \nvDash O$ and $A \vDash O$. Since the latter disjunct is logically impossible, this is equivalent to H making deductive observational assertions relative to A. Since the negation of an observation sentence is assumed to be observational, this is equivalent to H being falsifiable relative to A by claim 2.[17] □

Finally, falsifiability and relative falsifiability meet condition (D), for they explicitly allow auxiliary assumptions.

Only falsifiability and equivalent criteria fulfill condition (B) in the deductive case, and so it is good news that falsifiability can be phrased in terms of *modus ponens* and fulfills all other conditions of adequacy that pertain to criteria of deductive empirical significance. To meet condition of adequacy (H), any criterion of relative probabilistic empirical significance must therefore contain relative falsifiability as formulated in definition 8 as a special case. Then the corresponding absolute criterion also contains absolute falsifiability as a special case.

5 Elucidations of the conditions of adequacy for probabilistic criteria

The discussion of the adequacy of falsifiability is straightforward because concepts like the making of deductive assertions and deductive empirical equivalence are well-understood. On the other hand, in the discussion of Sober's interpretation of the inequality in the criterion of contrastive testability in §2 and the discussion of the condition of adequacy (B), it has already become apparent that dealing with undefined probabilities is not an entirely

[16] To allow sets of sentences and higher order logic, the definition must be phrased as "Two theories H_1 and H_2 are deductively empirically equivalent relative to assumptions A if and only if for every set Ω of observation sentences and every observation sentence O, $\Omega \cup H_1 \cup A \vDash O$ if and only if $\Omega \cup H_2 \cup A \vDash O$".

[17] The proof for sets of sentences is analogous, except for an additional existential quantification over sets of observation sentences.

Values of likelihoods		$\Pr(O \mid H_1 \wedge A) \neq \Pr(O \mid H_2 \wedge A)$								
$\Pr(O \mid H_1 \wedge A)$	$\Pr(O \mid H_2 \wedge A)$	1	2	3	4	5	6	7	8	9
x	y	$T\ldots$								
x	x	$F\ldots$								
x	U	T	T	T	F	F	F	U	U	U
U	U	T	F	U	T	F	U	T	F	U

Table 1: The nine possible interpretations of the inequality in definition 6 depending on the values of the probabilities, where x, y with $x \neq y$ are acceptable values for likelihoods, 'T' stands for 'true', 'F' stands for 'false', and 'U' stands for 'undefined'.

straightforward matter. And even though definition 4 treats the case of undefined likelihoods explicitly, it still involves some lacunae. Specifically, if one of the likelihoods is not defined, it is not obvious how to treat the inequality (3), and not entirely obvious how to treat the whole definition. This is because the definition is logically a conjunction with the inequality as a conjunct, and if the inequality is not defined, it is unclear whether a conjunction with one undefined conjunct is undefined as well.

It is uncontentious that $\Pr(S \mid H)$ is defined when H assigns a unique probability to S. But as Sober himself states when arguing for the need for auxiliary assumptions, theories alone often do not assert anything, and thus do not assign a probability to any observation sentence. And even with auxiliary assumptions, no theory will make assertions about everything.[18] Rather, assuming $H \wedge A$ restricts the probabilities that can be assigned to *some* sentences S to a subset of the interval $[0,1]$, possibly to one specific value $x \in [0,1]$, while the range of probabilities for some sentences will remain unrestricted. The conditional probability $\Pr(S \mid H \wedge A)$ can then either always be read as the set of probabilities that S can have under the assumption of $H \wedge A$, from $[0,1]$ to proper subsets thereof down to the singleton set $\{x\}, x \in [0,1]$. Or $\Pr(S \mid H \wedge A)$ may be read as defined only when it is a set of some specific kind considered acceptable (e. g., an interval or a singleton set), and undefined in all other cases.

The second reading of conditional probabilities leads to different interpretations of the inequality (3), depending on the treatment of formulas that contain undefined terms. Sober seems to assume the validity of classical logic, so that $\neg\varphi$ is false if and only if φ is true and tautologies are always true. This excludes some of the possible interpretations as given in table 1: Considering again the case where p and q are defined and different, while a is undefined, it is clear that interpretations 4–6 (5 being Sober's) are inconsistent

[18]For the technically inclined: For any even remotely plausible theory H, $H \wedge A$ can be observationally complete only for very restricted languages, where H is observationally complete if and only if for all observations O, $\Pr(O \mid H \wedge A) = x, x \in [0,1]$. It is thus always possible to expand the language by well-interpreted observation terms so that H fails to be observationally complete.

because they lead to $p = a$, $a = q$, and $p \neq q$. When a is undefined, interpretations 1, 3, 7, and 9 do not render $a \neq a$ false, and thus they are also excluded. Interpretation 8 is excluded if one demands that classical logic be truth-preserving and at least one disjunct of a true disjunction be true. For then, if p and q are defined and identical, while a and b are undefined, $p = q$ and $a = b$ are true, and entail $a = p \vee b \neq q$, which according to interpretation 8 has two undefined disjuncts. The remaining interpretation 2 can be seen as following from the introduction of the special value 'undefined' for probability-terms.

Under these assumptions, there are thus two possible interpretations of the inequality:

1. When all sets of probabilities are acceptable, the inequality is true if and only if the two sets differ. Otherwise, it is false.

2. When some sets of probabilities are unacceptable, the inequality is true if and only if its two sides are defined and different, or one side is defined and the other one is not. Otherwise, the inequality is false.

It is clear that the inequality is true more often for interpretation 1 than for interpretation 2, since in interpretation 1 it is true whenever a set of probabilities on one side differs from the set on the other side, but also when there is a difference between two sets that are unacceptable under the second reading of the likelihoods. It is also clear that Sober does not subscribe to interpretation 1, since in that case, there are no undefined likelihoods. In fact, he developed his concept of contrastive testability under the assumption that only singleton sets of probabilities are acceptable (Sober 2010, 3).

If all probability assignments are truth value-like, then the two interpretations of the inequality are equivalent, independently of whether $\{0, 1\}$ is an acceptable set of probabilities. For if $\{0, 1\}$ is an acceptable set, the interpretations are trivially equivalent; if $\{0, 1\}$ is not acceptable, the inequality is false if and only if both likelihoods have the values $\{0\}$, $\{1\}$, or $\{0, 1\}$/undefined. Otherwise, the inequality is true.

With these two readings of probability and the corresponding interpretations of the inequality, definition 4 is now indeed defined in all cases because the inequality is always either true or false. And it is also possible to explicate condition of adequacy (C) in line with Sober's position, for he states that "empirically equivalent theories have identical likelihoods" for any observation sentence O (Sober 1990, 399), which, treating the case of undefined likelihoods explicitly, leads directly to

Definition 10. Theories H_1 and H_2 are *probabilistically empirically equivalent relative to* auxiliary assumptions A if and only if for all observation sentences O,

(I) $\Pr(O \mid H_1 \wedge A)$ and $\Pr(O \mid H_2 \wedge A)$ are not defined or

(II) $\Pr(O \mid H_1 \wedge A)$ and $\Pr(O \mid H_2 \wedge A)$ are defined and $\Pr(O \mid H_1 \wedge A) = \Pr(O \mid H_2 \wedge A)$.

As defined, probabilistic empirical equivalence contains deductive empirical equivalence as a special case:

Claim 4. *Let H_1 and H_2 be deductive theories and let all probability assignments be truth value-like. Then H_1 and H_2 are probabilistically empirically equivalent relative to A if and only if H_1 and H_2 are deductively empirically equivalent relative to A.*

Proof. Since interpretation 1 and interpretation 2 are equivalent, it suffices to prove the claim for interpretation 1. H_1 is probabilistically empirically equivalent to H_2 relative to A if and only if for all observation sentence O, $H_1 \wedge A$ restricts the probability to the same set of values as $H_2 \wedge A$. Since for any H, $\Pr(S \mid H \wedge A) = 0$ if and only if $\Pr(\neg S \mid H \wedge A) = 1$, this is the case if and only if for every sentence, $H_1 \wedge A$ restricts the probability to 1 iff $H_2 \wedge A$ does. By claim 12 (see appendix), this holds if and only if $H_1 \wedge A$ and $H_2 \wedge A$ entail the same observation sentences, that is, if H_1 and H_2 are deductively empirically equivalent relative to A. □

Another reason to consider definition 10 a good explication of probabilistic empirical equivalence is that, if the probabilities of H_1 and H_2 given A are defined, it bears the same relation to the Bayesian criterion of empirical significance given in definition 2 as the criterion of deductive empirical equivalence bears to falsifiability: If two theories are deductively empirically equivalent, then either both or neither are deductively empirically significant (see appendix, claim 14). Analogously, the following holds:

Claim 5. *If $\Pr(H_1 \mid A)$ and $\Pr(H_2 \mid A)$ are defined and H_1 is probabilistically empirically equivalent to H_2 relative to auxiliary assumptions A, then, relative to A, observations are relevant for H_1 if and only if observations are relevant for H_2.*

Proof. If O is any observation sentence for which $\Pr(O \mid H_1 \wedge A) = \Pr(O \mid H_2 \wedge A)$, then $\Pr(O \mid H_1 \wedge A) = \Pr(O \mid A)$ if and only if $\Pr(O \mid H_2 \wedge A) = \Pr(O \mid A)$. Therefore $\Pr(H_1 \mid A) = \Pr(H_2 \mid O \wedge A)$ if and only if $\Pr(H_2 \mid A) = \Pr(H_2 \mid O \wedge A)$ (see appendix, claim 15). Thus observations are relevant for both H_1 and H_2 or for neither. □

Condition of adequacy (B) does not need to be fully explicated to show that it is not met by contrastive testability, and since it immediately provides a new criterion of empirical significance, it will be discussed in depth in §7.

6 Contrastive testability and the conditions of adequacy

Given the problems with the application of contrastive testability mentioned in §2, it should come as no surprise that contrastive testability fails to meet some of the conditions of adequacy. But it is perhaps not obvious which of the conditions it fails to meet and why it fails to meet them.

Condition of adequacy (A) is that a criterion of empirical significance must not be trivial, and claim 1 shows that definition 3 does not meet this condition. Definition 6 does, however: Choose $A = \varnothing$, two non-observational, non-equivalent sentences S and S', and, for some observation sentence O, $H_1 \Dashv\vDash S \land \Pr(O) = p$ and $H_2 \Dashv\vDash S' \land \Pr(O) = q$ for some probabilities p and q. Then H_1 and H_2 are never equivalent, and H_1 can be contrastively tested against H_2 if and only if $p \neq q$, so that many contingent theories can and many contingent theories cannot be tested against each other relative to A. Since $\varnothing$ is an honest set, H_1 and H_2 can also be absolutely tested against each other. As in the case of absolute falsifiability, it is impossible to prove that there are contingent theories that cannot be tested against each other relative to any honest set without more precise notions of justification and dependence.

Though non-trivial, contrastive testability fails to meet the two most important conditions of adequacy, conditions (B) and (C). That some theories that do not make probabilistic assertions and theories that are probabilistically empirically equivalent to tautologies are contrastively testable can be inferred from an example that Sober (1999, n. 24) attributes, in a different context, to Greg Mougin:[19]

> Let H_1 = God created the eye, E = Jones is pregnant, A = Jones is sexually active, and H_2 = Jones used birth control. It is possible to test H_1 against H_2; given independently attested background assumptions A, E favors H_1 over H_2.

In the example, the observation sentence E is assigned one probability by the background assumptions alone (since H_1 is not about Jones at all), and another by the conjunction of the background assumptions and H_2. Now choose $H_1 \Dashv\vDash \top$. Then H_1 does not make any assertions and hence no observable ones, and H_1 has trivially as much empirical content as a tautology, but it can still be contrastively tested against H_2, both relative to A and absolutely, since the justification of A does not depend on E.

Condition (D) is fulfilled by design because contrastive testability allows for auxiliary assumptions, and condition (E) is fulfilled because the theory H_2 about Jones's use of birth control is probabilistic and can be tested against H_1. Also by design, contrastive testability does not rely on prior probabilities or the likelihoods of the negation of theories and thus meets condition of adequacy (F).

Contrastive testability fails to meet condition (G) simply because so far, no Bayesian criterion of empirical significance has been suggested that is equivalent to contrastive testability when all occurring probabilities are defined. In this case, as can be seen from its logical structure, relative contrastive testability is not equivalent to the typical Bayesian criterion of empirical significance given by definition 2. Instead, it is almost the contradictory of probabilistic empirical equivalence:

[19]Unlike in the example by Salmon (1971, 29–88), it is this time not John Jones who is using birth control, but his wife.

Claim 6. *H_1 cannot be tested against H_2 relative to auxiliary assumptions A if and only if for all observation sentences O,*

(I) $\Pr(O \mid H_1 \wedge A)$ *or* $\Pr(O \mid H_2 \wedge A)$ *is not defined, or*

(II) $\Pr(O \mid H_1 \wedge A) = \Pr(O \mid H_2 \wedge A)$.

Thus contrastive testability differs from definition 10 in that it demands different relations between defined and undefined likelihoods.

In principle, a probabilistic two-place predicate may contain a deductive one-place predicate as a special case and thus meet condition (H). For example, if the probability assignments are truth value-like, '$\Pr(O \mid H_1) = .5 \vee \Pr(O \mid H_2) = 1$' is equivalent to '$H_2 \vDash O$' because the first argument, H_1, becomes irrelevant. But since contrastive testability is symmetric, either both or neither of its two arguments are irrelevant for truth value-like assignments and thus it cannot meet condition (H).

H_1 can be tested against H_2 if and only if their defined likelihoods differ for at least one observation sentence. If only singleton sets of probabilities are acceptable (as Sober assumes for his criterion), this means that at least with respect to one observation, one of the two theories must be wrong. Arguably, then, contrastive testability explicates what it means for two theories to be probabilistically empirically incompatible for the special case that only singleton sets of likelihoods are acceptable. This is borne out by the comparison with

Definition 11. Theories H_1 and H_2 are deductively empirically incompatible relative to auxiliary assumptions A if and only if there there is an observation sentence O such that $H_1 \wedge A \vDash O$ and $H_2 \wedge A \vDash \neg O$.

Claim 7. *Let H_1 and H_2 be deductive theories, let all probability assignments be truth value-like, and let the set $\{0, 1\}$ be unacceptable as a value of a likelihood. Then H_1 can be tested against H_2 relative to A if and only if H_1 and H_2 are deductively empirically incompatible relative to A.*

Proof. H_1 can be tested against H_2 if and only if there is an observation O such that the likelihood of one theory for O is 0, while the other one is 1. Without loss of generality, assume $\Pr(O \mid H_1 \wedge A) = 1$ and $\Pr(O \mid H_2 \wedge A) = 0$, that is, $\Pr(\neg O \mid H_2 \wedge A) = 1$. By claim 12 (see appendix), this holds if and only if $H_1 \wedge A \vDash O$ and $H_2 \wedge A \vDash \neg O$. □

Thus, if only singleton sets are acceptable as values of likelihoods, then contrastive testability contains as a special case a criterion for deductive theories that relies only on *modus ponens*. It is only the wrong one.

7 Explicating probabilistic empirical significance

Contrastive testability does not meet all criteria of adequacy, but that might just be because the criteria cannot all be met at once. I will argue that this is not so by suggesting a criterion of empirical significance that does meet all the conditions. First, however, I want to discuss an intuitively attractive but flawed criterion.

One may think of defining that a theory is not empirically significant if and only if it cannot be tested against *any* theory. That is, H_1 is not empirically significant if and only if for all suitable auxiliary assumptions A and all theories H_2, claim 6 is true. But this definition is inordinately inclusive. For assume that H_1 is not such that all assertions from suitable auxiliary assumptions become undefined, that is, $\Pr(O \mid A)$ and $\Pr(O \mid H_1 \wedge A)$ are defined (though possibly identical) for some O and some suitable A. Then H_1 is empirically significant if there is any H_2 such that $\Pr(O \mid H_2 \wedge A)$ is defined and different from $\Pr(O \mid A)$. For if $\Pr(O \mid A) = \Pr(O \mid H_1 \wedge A)$, H_1 can be tested against H_2, and if $\Pr(O \mid A) \neq \Pr(O \mid H_1 \wedge A)$, H_1 can be tested against any tautology. The premises of this argument are commonly true, for example, according to Sober, if H_1 is "God created the eye", O is "Jones is pregnant", and A is "Jones is sexually active", for then H_2 can be "Jones used birth control". Choosing $H_1 \models\!\mid \top$, the argument shows that tautologies are empirically significant, which runs afoul of conditions of adequacy (B) and (C). It is also straightforward to show that conditions (G) and (H) are not met.

A more promising path to a criterion of probabilistic significance leads through the criterion of probabilistic empirically equivalence, definition 10. It follows from claims 2 and 3 that a theory makes deductive observational assertions if and only if it is not deductively empirically equivalent to a tautology. This suggests conditions of adequacy (B) and (C) are in fact equivalent, so that a theory makes probabilistic observational assertions relative to auxiliary assumptions A if and only if it is not probabilistically empirically equivalent to a tautology relative to A according to definition 4. Because $\Pr(O \mid \top \wedge A) = \Pr(O \mid A)$, this leads to

Definition 12. Theory H *makes probabilistic observational assertions relative to* A if and only if there exists an observation sentence O such that

(I) $\Pr(O \mid H \wedge A)$ is defined if and only if $\Pr(O \mid A)$ is not defined or

(II) $\Pr(O \mid H \wedge A)$ and $\Pr(O \mid A)$ are defined and $\Pr(O \mid H \wedge A) \neq \Pr(O \mid A)$.

With the help of definition 5 of an honest set, one can give

Definition 13. Theory H *makes probabilistic observational assertions* if and only if H makes probabilistic observational assertions relative to an honest set A.

It may be considered problematic that a theory H makes assertions if assuming H makes it impossible to assign a probability to an observation that otherwise would be assigned a probability by the auxiliary assumptions alone. To render such theories empirically non-significant, the biconditional in condition (I) of definition 12 could be made into a conjunction. But it is plausible that a theory that asserts that some observable regularity breaks down does make assertions. It certainly is pragmatically relevant when some observation sentence can, contrary to the auxiliary assumptions, not be assigned a probability.

Definitions 12 and 13 rely only on concepts that Sober uses himself, and should therefore be conceptually unproblematic for him. He does not discuss the domains of applicability of the concepts, but with one exception, the domains can just be assumed to be the same for definitions 12 and 13 as they are for contrastive testability. The exception is the term $\Pr(O \,|\, A)$. Sober could argue that the concept of a likelihood cannot be applied to tautologies because the auxiliary assumptions themselves assign probabilities to no or too few observation sentences. Sober (2008, 29f) in fact shortly discusses $\Pr(O)$, but not in connection with auxiliary assumptions. The discussion therefore does clearly not apply to definition 12, and it does not apply to definition 13 because H is testable if there are *some*, not necessarily tautological, suitable auxiliary assumptions A such that $\Pr(O \,|\, H \wedge A)$ differs from $\Pr(O \,|\, A)$. Sober's original definition 3 and definition 5 of an honest set put no restrictions on individual elements of A except that they be justified (in Sober's definition, independently of a specific observation sentence). Therefore whole theories can be included in the background assumptions. Since Sober introduces auxiliary assumptions to allow for actual scientific practice, and assertions made by scientific theories in fact often use other scientific theories as auxiliary assumptions, such an inclusion obeys letter and spirit of Sober's criterion. Since scientific theories H are supposed to make observational assertions, $\Pr(O \,|\, H \wedge A)$ will often be defined. And the inclusion of H into the auxiliary assumptions is then just the notational change to $\Pr(O \,|\, A')$ with $A' \mathrel{\models\!\!\!\dashv} H \wedge A$.

It now follows from condition of adequacy (B) that all and only theories that fulfill definitions 12 or 13 are probabilistically empirically significant. Without any basic conceptual problems, a theory can then be defined to be empirically significant (relative to A) if and only if it it makes observational assertions (relative to A), that is, if and only if it makes probabilistic observational assertions (relative to A) or it is falsifiable (relative to A).

As argued in §4.2, falsifiability fulfills all appropriate conditions of adequacy. I now want to show that this new definition of probabilistic empirical significance does, too. That it is non-trivial and thus meets condition of adequacy (A) is easily shown since sentences without observational terms are not testable relative to $\varnothing$. And in the example with Jones's pregnancy, the theory that Jones uses birth control (H_2) has a different likelihood given the auxiliary assumption A that Jones is sexually active than A alone, and therefore H_2 is testable relative to A. This example also shows that there are positive instances of absolute testability. As in the case of contrastive testability, a proof that there are also negative instances of testability would presume more precise conceptions of justification and dependence. That restrictions (III) and (IV) of definition 3 are not sufficient to avoid a trivialization of testability can be shown by choosing $H_2 \Dashv\vDash \top$ in claim 1. In this case, the restriction to honest sets also precludes any trivialization that restriction (IVa) could preclude, for if an element of A depends on H, $H \in A$, so that $\Pr(O\,|\,H \wedge A) = \Pr(O\,|\,A)$ for all O.

Definitions 12 and 13 trivially fulfill condition of adequacy (B). With definition 10 as criterion of probabilistic empirical equivalence, testability meets condition (C) as well. Since condition (C) only states that empirical equivalence to a tautology is a sufficient condition for empirically non-significance, (C) is also met if the biconditional of condition (I) in definition 12 is substituted by a conjunction, so that more theories are empirically non-significant.

Since definitions 12 and 13 allow the use of auxiliary assumptions, they meet condition of adequacy (D). Condition (E) is met because the theory about Jones's use of birth control is probabilistic and makes probabilistic observational assertions. By design, definitions 12 and 13 fulfill condition of adequacy (F).

Condition of adequacy (G) is met because of

Claim 8. *If all occurring probabilities are defined, then H makes probabilistic observational assertions if and only if observations are relevant for H.*

Proof. For all observations O, $\Pr(O\,|\,H \wedge A) \neq \Pr(O\,|\,A)$ if and only if $\Pr(H\,|\,O \wedge A) \neq \Pr(H\,|\,A)$ (see appendix, claim 15). Therefore there is an O such that $\Pr(O\,|\,H \wedge A) \neq \Pr(O\,|\,A)$ if and only if there is an O such that $\Pr(H\,|\,O \wedge A) \neq \Pr(H\,|\,A)$. □

Definition 12 fulfills condition of adequacy (H), because it generalizes classical falsifiability:

Claim 9. *Let H be a deductive theory, let all probability assignments be truth value-like, and let the negation of an observation sentence again be an observation sentence. Then H makes probabilistic observational assertions relative to A if and only if H is falsifiable relative to A.*

Proof. Since interpretation 1 and interpretation 2 of the inequality in condition (II) are equivalent, it suffices to prove the claim for interpretation 1. H does not make probabilistic assertions relative to A if and only if for all observation sentence O, $H \wedge A$ restricts the probability to the same set of values as A. This is the case if and only if for every sentence, $H \wedge A$ restricts the probability to 1 iff A does. By claim 12 (see appendix), this holds if and only if $H \wedge A$ and A entail the same observation sentences, that is, if and only if H makes no deductive observational assertions relative to A. By claim 2, this holds if and only if H is not falsifiable relative to A. □

Note that the proof also holds if the biconditional of condition (I) in definition 12 is substituted by a conjunction, because then H makes no observational assertions if and only if for all O, $H \wedge A$ restricts the probabilities to the same set of values as A, or A restricts the probabilities more than $H \wedge A$. But the latter is impossible since A restricts the probabilities of an observation sentence to $\{0\}$ or $\{1\}$ only if $H \wedge A$ does. If the suitable auxiliary assumptions for falsifiability are given by honest sets, condition (H) is also fulfilled by definition 13.

Therefore, definition 12 and arguably definition 13 fulfill all conditions of adequacy that Sober wants a criterion of empirical significance to meet. Additionally, they also make it possible to evaluate one theory, ID for example, independent of another one like ET. Specifically, ET might be empirically significant without ID being so. Finally, Sober's use of 'testability' as a one-place predicate is not only meaningful, but also correct when interpreted by the new definition: 'Undetectable angels exist' arguably makes no observational assertions relative to any honest set of sentences, and 'This coin has a probability of .5 of landing heads each time it is tossed' makes observational assertions, for it assigns a probability of .5 to an observation sentence relative to $\varnothing$.

8 Conclusion

The argument against contrastive testability as a criterion of probabilistic empirical significance can be summarized as follows: First, its restrictions on auxiliary assumptions are in part unjustified, in part misleading, and so weak that they render the criterion trivial. Second, even avoiding the question of suitable auxiliary assumptions, the criterion fails to meet four criteria of adequacy that follow from Sober's position and the intended application of contrastive testability: It does not exclude all theories that make no observational assertions, nor all theories that are empirically equivalent to tautologies. It is not equivalent to a Bayesian criterion of empirical significance when all probabilities are defined, and it does not contain falsifiability as a special case. This last property is important because a criterion of probabilistic empirical significance should contain *some* adequate criterion of deductive empirical significance as a special case, and falsifiability is the only adequate criterion.

However, Sober's search for a criterion of probabilistic empirical significance is no failure. For one, his defense of the assumptions underlying the search is still relevant. Furthermore, if only unique probabilities are acceptable as likelihoods, then contrastive testability is a plausible explication of probabilistic empirical incompatibility.

Given that contrastive testability is not an adequate criterion of empirical significance, I have suggested to consider a theory empirically significant relative to a set of auxiliary assumption if and only if it is not deductively or probabilistically empirically equivalent to a tautology relative to the auxiliary assumptions. This definition fulfills all eight conditions of adequacy. To arrive at an absolute criterion of empirical significance, I have tentatively suggested the notion of an honest set. Its use blocks some known trivialization proofs, but since the definition relies on an intuitive understanding of justification, further refinements of the absolute criterion of empirical significance will depend on an adequate explication of justification.

A Additional proofs

Claim 10. *If $S \vDash O$, then*

$$\Pr\big(O \mid H_2 \wedge (\neg H_1 \vee S)\big) \neq 1 \Longleftrightarrow \Pr(O \mid H_2 \wedge \neg H_1) \neq 1\,, \tag{4}$$

assuming $\Pr\big(H_2 \wedge (\neg H_1 \vee S)\big) \neq 0$ *and* $\Pr\big(H_2 \wedge \neg H_1\big) \neq 0$.

Proof. Since $S \vDash O$, $\Pr(O \wedge S \wedge Q) = \Pr(S \wedge Q)$ for any Q. Thus

$$\begin{aligned}\Pr\big(O \mid H_2 \wedge (\neg H_1 \vee S)\big) &= \frac{\Pr(O \wedge H_2 \wedge \neg H_1) + \Pr(O \wedge H_2 \wedge S) - \Pr(O \wedge H_2 \wedge \neg H_1 \wedge S)}{\Pr(H_2 \wedge \neg H_1) + \Pr(H_2 \wedge S) - \Pr(H_2 \wedge \neg H_1 \wedge S)} \\ &= \frac{\Pr(O \wedge H_2 \wedge \neg H_1) + c}{\Pr(H_2 \wedge \neg H_1) + c}\end{aligned} \tag{5}$$

The claim follows immediately. □

Claim 11. *For every language $\mathscr{L}$ and every $\mathfrak{M}$,* $\Pr_{\mathfrak{M}} : \mathfrak{P}\mathscr{L} \to \{0,1\}, \Pr_{\mathfrak{M}}(\Sigma) = 1 \Longleftrightarrow \mathfrak{M} \vDash \Sigma$ *is a probability assignment.*

Proof. Show that for all $\Sigma, \Xi \in \mathscr{L}$ and any $\mathfrak{M}$ it holds:

1. $\Pr_{\mathfrak{M}}(\Sigma) \geq 0$,
2. $\Pr_{\mathfrak{M}}(\{\top\}) = 1$, and
3. if Σ and Ξ are finite and $\Sigma \cup \Xi \vDash \bot$, $\Pr_{\mathfrak{M}}(\{\bigwedge \Sigma \vee \bigwedge \Xi\}) = \Pr_{\mathfrak{M}}(\Sigma) + \Pr_{\mathfrak{M}}(\Xi)$.

1 and 2 are immediate. 3 holds because for $\Sigma \cup \Xi \vDash \bot$, $\mathfrak{M} \nvDash \Sigma$ or $\mathfrak{M} \nvDash \Xi$, so that $\Pr_{\mathfrak{M}}(\bigwedge \Sigma \vee \bigwedge \Xi) = 1$ if and only if either $\mathfrak{M} \vDash \Sigma$ or $\mathfrak{M} \vDash \Xi$ but not both, which holds if and only if $\Pr_{\mathfrak{M}}(\Sigma) = 1$ or $\Pr_{\mathfrak{M}}(\Xi) = 1$ but not both, that is, $\Pr_{\mathfrak{M}}(\Sigma) + \Pr_{\mathfrak{M}}(\Xi) = 1$ □

Claim 12. *For any sets $\Sigma, \Xi \subseteq \mathscr{L}$ of sentences, $\Sigma \vDash \Xi$ if and only if for all $\mathfrak{M}$ it holds: If* $\Pr_{\mathfrak{M}}(\Sigma) \neq 0$ *then* $\Pr_{\mathfrak{M}}(\Xi \mid \Sigma) = 1$.

Proof.

$$\begin{aligned}\Sigma \vDash \Xi &\Longleftrightarrow \forall \mathfrak{M}\big[\mathfrak{M} \vDash \Sigma \Rightarrow \mathfrak{M} \vDash \Xi\big] \\ &\Longleftrightarrow \forall \mathfrak{M}\big[\Pr_{\mathfrak{M}}(\Sigma) = 1 \Rightarrow \Pr_{\mathfrak{M}}(\Xi) = 1\big] \\ &\Longleftrightarrow \forall \mathfrak{M}\left[\Pr_{\mathfrak{M}}(\Sigma) \neq 0 \Rightarrow \Pr_{\mathfrak{M}}(\Xi \mid \Sigma) = \frac{\Pr_{\mathfrak{M}}(\Xi)}{\Pr_{\mathfrak{M}}(\Xi \cup \Sigma)} = 1\right]\end{aligned}$$

□

Claim 13. *If the negation of an observational sentence is observational, the following holds:* $H \cup A \nvDash \bot$ *and there is a set* Ω *of observation sentences such that* $\Omega \cup A \nvDash \bot$ *and* $\Omega \cup H \cup A \vDash \bot$ *if and only if there are a set* Ω *of observation sentences and an observation sentence* O *such that* $\Omega \cup H \cup A \vDash O$ *and* $\Omega \cup A \nvDash O$.

Proof. '⇒': If $\Omega \cup H \cup A \vDash \bot$, then $\Omega \cup H \cup A \vDash O$ for any observation sentence O. Since $\Omega \cup A \nvDash \bot$, there is some O such that $\Omega \cup A \nvDash O$.

'⇒': For O and Ω with $\Omega \cup H \cup A \vDash O$ and $\Omega \cup A \nvDash O$, $\Omega \cup \{\neg O\} \cup A \nvDash \bot$ and $\Omega \cup \{\neg O\} \cup H \cup A \vDash \bot$. □

Claim 14. *If the negation of an observation sentence is again an observation sentence, and* H_1 *is deductively empirically equivalent to* H_2 *relative to* A*, then, relative to* A*,* H_1 *is falsifiable if and only if* H_2 *is falsifiable.*

Proof. Assume that for all observation sentences O and sets of observation sentence Ω, $H_1 \cup \Omega \cup A \vDash O$ if and only if $H_2 \cup \Omega \cup A \vDash O$. Then, for all Ω and O, $H_1 \cup \Omega \cup A \vDash O$ and $\Omega \cup A \nvDash O$ if and only if $H_1 \cup \Omega \cup A \vDash O$ and $\Omega \cup A \nvDash O$. Thus there are Ω and O such that $H_1 \cup \Omega \cup A \vDash O$ and $\Omega \cup A \nvDash O$ if and only if there are Ω and O such that $H_1 \cup \Omega \cup A \vDash O$ and $\Omega \cup A \nvDash O$. By claim 2, this means that H_1 is falsifiable relative to A if and only if H_2 is falsifiable relative to A. □

Claim 15. *If* $\Pr(H \mid A)$ *is defined, then* $\Pr(H \mid O \wedge A) = \Pr(H \mid A)$ *if and only if* $\Pr(O \mid H \wedge A) = \Pr(O \mid A)$.

Proof. From

$$\frac{\Pr(H \mid O \wedge A)}{\Pr(H \mid A)} = \frac{\Pr(O \mid H \wedge A)}{\Pr(O \mid A)}, \tag{6}$$

the claim follows immediately. □

References

Ayer, A. J. (1936). *Language, Truth and Logic*. Victor Gollanz, London, 1st edition. Page numbers refer to the second edition (Ayer 1946). 16

Ayer, A. J. (1946). *Language, Truth and Logic*. Victor Gollanz, London, 2nd edition. 33

Belnap, N. (1993). On rigorous definitions. *Philosophical Studies*, 72:115–146. 6

Carnap, R. (1937). Testability and meaning—continued. *Philosophy of Science*, 4(1):2–35. 21

Carnap, R. (1950). *Logical Foundations of Probability*. University of Chicago Press, Chicago. References are to the 2nd edition (Carnap 1962). 2, 8

Carnap, R. (1956). The methodological character of theoretical concepts. In Feigl, H. and Scriven, M., editors, *The Foundations of Science and the Concepts of Psychology and Psychoanalysis*, volume 1 of *Minnesota Studies in the Philosophy of Science*. University of Minnesota Press, Minneapolis, MN. 21

Carnap, R. (1962). *Logical Foundations of Probability*. University of Chicago Press, Chicago, 2nd edition. 34

Carnap, R. (1963). Replies and systematic expositions. In Schilpp (1963), pages 859–1016. 3

Diamond, M. L. and Litzenburg, T. V., editors (1975). *The Logic of God: Theology and Verification*. Bobbs-Merill, Indianapolis, IN. 34

Duhem, P. (1914). *La théorie physique, son objet et sa structure*. Marcel Rivière, Paris, 2nd edition. References are to the translation (Duhem 1954). 17

Duhem, P. (1954). *The Aim and Structure of Physical Theory*. Princeton University Press, Princeton, NJ. 34

Feller, W. (1971). *An Introduction to Probability Theory and Its Applications*, volume 2. John Wiley & Sons, Inc., New York, N.Y., 2nd edition. 15

Flew, A. (1950). Theology and falsification. *University*, 1:1–8. References are to the reprint (Diamond and Litzenburg 1975, 257–259). 17

Frege, G. (1918). Der Gedanke: eine logische Untersuchung. *Beiträge zur Philosophie des deutschen Idealismus*, 1:58–77. References are to the translation (Frege 1956). 8

Frege, G. (1956). The thought: A logical inquiry. *Mind*, 65(259):289–311. 34

Friedman, M. (1982). Review. *The Journal of Philosophy*, 79(5):274–283. Review of (Van Fraassen 1980). 21

Hempel, C. G. (1952). *Fundamentals of Concept Formation in Empirical Sciences*, volume II,7 of *Foundations of the Unity of Science. Toward an International Encyclopedia of Unified Science*. The University of Chicago Press, Chicago and London. References are to the two volume edition. 2, 8

Hempel, C. G. (1965). Empiricist criteria of cognitive significance: Problems and changes. In *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*, pages 81–119. The Free Press, New York. 8

Justus, J. (2010). Evidentiary inference in evolutionary biology. *Biology and Philosophy*, forthcoming. Review of (Sober 2008). 2, 8

Kuipers, T. A. F. (2007). Introduction. Explication in philosophy of science. In Kuipers, T. A. F., editor, *General Philosophy of Science—Focal Issues*, volume 1 of *Handbook of the Philosophy of Science*, pages vii–xxiii. Elsevier, Amsterdam. 2

Lakatos, I. (1974). Falsification and the methodology of scientific research programmes. In Lakatos, I. and Musgrave, A., editors, *Criticism and the Growth of Knowledge*, volume 4 of *Proceedings of the International Colloquium in the Philosophy of Science*, pages 91–196. Cambridge University Press, Cambridge. 4

Lewis, D. (1988). Statements partly about observation. *Philosophical Papers*, 17:1–31. References are to the reprint (Lewis 1998, 125–155). 21

Lewis, D. (1998). *Papers in Philosophical Logic*. Cambridge Studies in Philosophy. Cambridge University Press, Cambridge. 35

Lutz, S. (2010a). Criteria of empirical significance: A success story. Forthcoming. 8, 13, 14, 20

Lutz, S. (2010b). On Sober on intelligent design. Forthcoming. 17

Popper, K. (1963). The demarcation between science and metaphysics. In Schilpp (1963), pages 183–226. 3

Popper, K. R. (1935). *Logik der Forschung: Zur Erkenntnistheorie der modernen Naturwissenschaft*, volume 9 of *Schriften zur wissenschaftlichen Weltauffassung*. Julius Springer, Wien. Published in 1934, nominal publication date incorrect. References are to the extended and revised translation (Popper 2000). 3, 8, 15, 35

Popper, K. R. (2000). *The Logic of Scientific Discovery*. Routledge, London and New York, NY. Translated and extended from the German original (Popper 1935) by Karl Popper, Julius Freed, and Lan Freed. 35

Psillos, S. (2000). Rudolf Carnap's 'Theoretical concepts in science'. *Studies in History and Philosophy of Science Part A*, 31:151–172. 21

Salmon, W. C. (1967). *The Foundations of Scientific Inference*. University of Pittsburgh Press, Pittsburgh, PA. 19

Salmon, W. C. (1971). *Statistical Explanation and Statistical Relevance*. University of Pittsburgh Press, Pittsburgh, PA. With contributions by Richard C. Jeffrey and James G. Greeno. 26

Schilpp, P. A., editor (1963). *The Philosophy of Rudolf Carnap*, volume 11 of *The Library of Living Philosophers*. Open Court Publishing Company, Chicago and LaSalle, IL. 34, 35

Smolin, L. (2006). *The Trouble With Physics: The Rise of String Theory, the Fall of a Science, and What Comes Next*. Houghton Mifflin, Boston, MA. 2

Soames, S. (2003). *The Dawn of Analysis*, volume 1 of *Philosophical Analysis in the Twentieth Century*. Princeton University Press, Princeton, NJ. 2

Sober, E. (1990). Contrastive empiricism. In Savage, C. W., editor, *Scientific Theories*, volume 14 of *Minnesota Studies in the Philosophy of Science*, pages 392–410. University of Minnesota Press, Minneapolis, MN. 2, 18, 24

Sober, E. (1999). Testability. *Proceedings and Addresses of the American Philosophical Association*, 73(2):47–76. 2, 8, 9, 11, 13, 16, 17, 18, 20, 26

Sober, E. (2002). Intelligent design and probability reasoning. *International Journal for Philosophy of Religion*, 52(2):65–80. 4, 18

Sober, E. (2007). What is wrong with intelligent design? *The Quarterly Review of Biology*, 82(1):3–8. 2, 3, 9, 10, 17, 18

Sober, E. (2008). *Evidence and Evolution: The Logic Behind the Science*. Cambridge University Press, Cambridge. 2, 3, 4, 5, 6, 7, 8, 9, 11, 16, 17, 18, 21, 29, 35

Sober, E. (2010). Comments on Sebastian Lutz's "On Sober's criterion of contrastive testability". Unpublished typescript. 2, 3, 7, 24

van Fraassen, B. C. (1980). *The Scientific Image.* The Clarendon Library of Logic and Philosophy. Clarendon Press, Oxford. 34

Woit, P. (2006). *Not Even Wrong: The Failure of String Theory and the Search for Unity in Physical Law*. Basic Books, New York. 2

Typeset in Bitstream Charter with pdfLaTeX, mathdesign, and microtype.