Testability and Candor: In Memory of Robert Nozick

Author(s): Sherrilyn Roush

Source: *Synthese*, Vol. 145, No. 2, Candor in Science (Jun., 2005), pp. 233-275
Published by: Springer
Stable URL: https://www.jstor.org/stable/20118592
Accessed: 27-06-2020 08:43 UTC

SHERRILYN ROUSH

TESTABILITY AND CANDOR

In Memory of Robert Nozick

ABSTRACT. On analogy with testimony, I define a notion of a scientific theory's lacking or having candor, in a testing situation, according to whether the theory under test is probabilistically relevant to the processes in the test procedures, and thereby to the reliability of test outcomes. I argue that this property identifies what is distinctive about those theories that Karl Popper denounced as exhibiting "reinforced dogmatism" through their self-protective behavior (e.g., psychoanalysis, Hegelianism, Marxism). I explore whether lack of candor interferes with the testing of theories, and conclude that (1) our default attitude toward theories that lack candor in a given test should be suspicion, but (2) the circumstance that a theory lacks candor in a testing situation does not preclude obtaining independent evidence for the auxiliary assumptions to which the theory is probabilistically relevant, and thereby eliminating the problem that lack of candor creates. Thus, Popper was right to think that lack of candor is a bad thing, but wrong to conclude that candor is a criterion of the scientificity of a theory. Seeing this requires recognition of some differences between intuitive relevance and probabilistic relevance, and proper appreciation of the notion of *screening off* and of the fact that probabilistic relevance is not transitive.

Sometimes saying that a person was not candid with us is simply a muted way of saying that they were lying. However, uttering falsehoods is not necessary for failing to be candid; in the most common usage, one fails to be candid when one fails to utter a relevant truth that one knows, especially a truth about one's own shortcomings unknown to someone who may be adversely affected by them. One may flout candor if what one does say, even if true, positively diverts attention from salient truths or from the fact that they are being omitted. To be candid, it is not enough that the things one *says* are true. When we believe that someone has been candid with us, what we believe is that if there were truths important to us, that the person knows and that we should know, this person would have told them to us, or at least would not have taken steps to hide them. A person is candid if she has the disposition to say the (relevant)

Synthese (2005) 145: 233–275
DOI 10.1007/s11229-005-3748-1

things she knows irrespective of whether their being known confers an immediate disadvantage on her. A person who is not candid with us often behaves so in order to protect herself, and possibly us, from the awkward consequences of the truth.

I will not try to press this further into a philosophical analysis of the concept of candor in human life. Rather, I want to use the intuitive notion of a person's being candid to call attention to some special problems about the testability of scientific theories. Most of us have given up the old project of defining testability as the property that makes scientific theories meaningful, and whose lack makes metaphysical propositions nonsensical. But regardless of its relation to meaning and language, the question of what makes a theory testable against experience remains a central issue for philosophy of science because of our interest in epistemic questions about how theories are justified. In this paper I will be concerned with stumbling blocks to testability that some theories face in virtue of what they claim about the world, and whether and how it is possible to overcome them. The type of problem I have in mind is something that caught the attention of Karl Popper when he described Hegelianism, Marxism, psychoanalysis, and philosophies of meaning as displaying "reinforced dogmatism," a phenomenon that he regarded as similar to "total ideology." This phenomenon, he thought, undermined the possibility of rational discussion, and represented a form of dishonesty.

While Popper did not give a precise characterization of this phenomenon, as far as I know, his comments on particular cases suggest that he thought these theories were not unfalsifiable because of a lack of empirical consequences, but rather because they had features that contributed in a different way to the theories' imperviousness to criticism. Consider the theory of psychoanalysis in what I will call the 'cartoon' version. For this theory, the problem is not whether it leads to observable claims; it does. The problem, for example, is that according to the theory, a patient's denial of an analyst's accounts of her psyche is an *indicator* of repression, since a repressed patient is likely to resist admission of the repressed material. But it is conceivable that the patient denies an account of herself not because of repression, but instead because the account of her as repressed is wrong and she knows it. Her denial would ordinarily be a potential falsifier of the account, but here according to the theory under scrutiny, this whole class of potential falsifiers—the patient's own testimony that an analyst's proposals about her repressions are incorrect—could have their critical power defused because the theory offers an explanation of them that is flattering to the theory's point of view. It seems that we would have to determine whether the theory was correct before we could fairly determine whether a certain type of evidence counts as a falsifier of that theory or not, and that seems to put the cart before the horse if what we want is empirical testing.[1]

What the theories Popper hated most have in common is—to put it metaphorically—that they have the capacity to protect *themselves* from empirical falsification, usually by means of what these theories claim about the way the world is. If we thought of a theory as a person, and asked what this self-protective behavior was analogous to, I think the closest notion would be a lack of candor. It is true that in testing theories what we are worried about is whether what they claim is true, not whether they tell us everything we want to know. It may thus seem that the question we should ask about theories if we are comparing them to human beings is whether they are "lying" outright, not whether they are candid. But though we are indeed concerned with whether theories (or their empirical consequences) are true, the apparent disanalogy between theories and candor just cited is due to the fact that our standard procedures for evaluating theories and people are slightly different. With people we generally believe their testimonies unless there is an indication, even a small indication, that they are lying, incompetent, or joking. Fortunately we do not generally believe what theories say just because they say it. The standard procedure is to test the theories. So the analogous questions arise about testimonies of people and tests of theories: 'Was that person's testimony candid?' is analogous to the question, 'Did anything about the theory (beyond the possibility that it makes no empirical difference) prevent that test of it from being revealing?' The idea is that the person, or theory, under scrutiny is making a contribution to the testimony, or test, by means of which we are trying to evaluate it, a contribution that could interfere with the worth of the testimony or test. The analogy is loose, and intended only as a heuristic: if it is not helpful, then think of 'lack of candor' as a placeholder for the notion that I will define precisely below.

Popper thought the self-protective aspect he saw in some theories was a clear indication that they were not scientific, because they did not lay themselves open to rational criticism. I will not assume that self-protectiveness renders a theory unscientific, but rather develop a view about why theories have the potential for the self-protection of which Popper so disapproved (and for other types of epistemic self-activity that he did not notice). It will then become clear that bona fide scientific theories may have this feature. This will show that lack of potential for self-protectiveness is not the straightforward criterion of scientificity for theories that Popper thought it was.

Allowing that proper science contains self-protective theories should not come as a great surprise, since many of Popper's critics—Thomas Kuhn, Imre Lakatos, and Paul Feyerabend, for instance—have emphasized through their notions of paradigm, research programme, and opportunism that what we might strictly consider dogmatism is not foreign, and may be essential, to successful science (Lakatos 1978; Feyerabend 1993; Kuhn 1996). But I will argue that there are ways of overcoming the epistemic problems that self-protective theories present, implying that such theories can be used to good effect in proper science, even in tests in which they are self-protective. This will show that it is not that truly scientific theories must lack the potential for self-protection, but that in good science a theory's potential for self-protection is well-handled.

I wish to draw a sharp distinction that these and other authors do not always adhere to between the dogmatic psychological dispositions proponents of theories may have and the potentially self-protective features of theories themselves. The two kinds of dogmatism often have similar consequences, but their remedies are entirely different. In particular, if, as I will put it, a theory *lacks candor*, then no degree of open-mindedness of disposition on the part of an adherent is alone likely to find, much less to solve, a problem that this may create. (Conversely, if a person is sufficiently dogmatic psychologically, then our overcoming the problems that may be created by lack of candor in her theory may not make any difference to her beliefs.) In order to identify and understand what the content of a theory can contribute to difficulties in testing it, I am going to assume in this paper that the human handlers of a theory have the best possible intentions and behaviors. This will focus attention on problems that would exist regardless of scientists' virtue. So, for example, the problems with testing that I will be interested in will never take the form of a scientist accepting an auxiliary hypothesis just because it is the only way she can think of to protect her pet theory from disconfirmation.

1. WHEN DOES A THEORY LACK CANDOR?

A test of a theory can fail to be revealing in many different ways. The accuracy of the instruments may be low, in measurement or otherwise. The test may yield false negatives or false positives. The test may scrutinize only an isolated part of a theory. The test may simply be botched. I will regard a theory as *lacking candor* only if the theory itself contributes to the test it undergoes being in some way nonrevelatory. A theory will contribute to any testing problem concerning it in the trivial sense that it is the theory's claims that are under scrutiny. To lack candor, in the intuitive sense, the theory will have to make an interfering contribution over and above this. The examples Popper was worried about make a contribution of the following form: the theory is probabilistically relevant to an auxiliary assumption about an instrument, process, or method that produces evidence used in the testing of the theory.

A statement, Y, is *probabilistically relevant* to another, X, when the probability of X given Y is not equal to the probability of X, i.e., $\mathrm{P}(X/Y) \neq \mathrm{P}(X)$. Intuitively, Y's being so makes a difference to how probable it is that X is so. Y may be either *positively* or *negatively* relevant depending on whether $\mathrm{P}(X/Y) > P(X)$ or $\mathrm{P}(X/Y) < P(X)$. This notion of relevance is distinct from, weaker than, and therefore fulfilled in a wider variety of cases than, logical relevance; Y and X are *logically relevant* just in case either Y proves X or refutes X or X proves Y or refutes Y. From here on, where I use the term 'relevance' alone, I will mean probabilistic relevance.[2] When two statements fail to be probabilistically relevant they are said to be probabilistically independent. That is, X is *probabilistically independent* of Y when and only when $\mathrm{P}(X/Y) = P(X)$. Intuitively, whether Y is true makes no difference to the probability of X. In conformity with the fact that probabilistic relevance is weaker than logical relevance, probabilistic independence is stronger than logical independence.

The condition I have stated in terms of probabilistic relevance is necessary for a theory to lack candor in the manner that I think Popper's examples do, but in §2 I will combine it with another constraint in order to obtain a set of necessary and sufficient conditions. I leave open the possibility that there are other ways than this for a theory to lack candor in the intuitive sense of protecting itself from testing or interfering with its testing. What I am defining now is only a notion that I think captures what is distinctive in Popper's examples. Whether or under what conditions lack of candor is a problem and, when a problem, how it can be overcome, are questions I address in subsequent sections of this paper.[3]

This necessary condition for lack of candor describes something that the cases Popper was exercised about have in common. (His cases also fulfill the refined condition of the next section.) The problem with psychoanalytic theory described above fulfills the condition exactly. The patient's testimony is one method for testing an account of a patient's psyche (and thereby, to some degree, a theory of human psyches in general). However, according to psychoanalysis (in the cartoon version I have described), what the patient professes may be discounted, under the rubric of 'denial,' if what she professes is a rejection of the analyst's theory. According to an opponent of that theory, the patient's own testimony, taken at face value, might be considered some of the most important evidence against the theory. The theory is probabilistically relevant to this method of testing it, because what is happening in a testifying patient, what helps to produce his testimony, is something about which the theory happens to make claims. This probabilistic relevance makes it hard to imagine the patient's testimony as evidence whose import for this theory could be judged from any unbiased point of view.

The caricature of Hegelianism that Popper commented on has a similar structure. The Hegelian supposedly thinks that there are contradictions in the nature of things, i.e., Hegel overcame Kant's antinomies by creating a system in which contradictions are tolerated. More charitably, we may say that Hegel thought contradictions had to be tolerated in a true view of reality because reality was evolving. Even if the Hegelian does not think that granting the truth of one (or four) contradictions is committing oneself to all contradictions (and every other statement), his opponent will usually think this, because in the usual logic a pair of contradictory propositions implies every proposition. Thus if we believe, as Popper did, that "any criticism of any theory whatsoever, must be based on the method of pointing out some sort of contradiction, either within the theory itself or between the theory and some facts, ..." then Hegel's theory, which implies (according to Popper) that all contradictions are to be tolerated, defuses all possible criticism of itself[4] (Popper 1989, 327; Popper 1966, 215). That is, if you believe the theory then you have grounds for regarding what are normally criticisms of your theory as unproblematic. Intuitively, this does not seem to be a candid theory since it protects itself from criticism. It is also fair to say that something like the condition above holds: Hegel's theory is relevant to, has something to say about, the method of testing that theory.

An important difference between these two cases is that if Popper's claim about all criticism being based on contradiction is right, then Hegel's theory protects itself against *all possible* criticism, whereas nothing so global follows from the analogous admission about Freud's theory. A patient's direct testimony about a specific claim of the analyst is only one of the possible ways of probing the claims psychoanalytic theory makes about relations between, for example, repression and pathology. Adolf Grünbaum's careful studies have argued persuasively that psychoanalysis is testable, indeed falsifiable, and that actual empirical tests have cast some doubt on it (Grünbaum 1984, 97–126). So even if we suppose with Popper that psychoanalytic theory cannot be tested properly via direct patient testimony about possibly repressed matters, it may be testable by other means, and apparently is. A tendency to self-protection in some test does not imply that a theory is untestable in general. Therefore, lack of candor should be defined in such a way that it is a local property, a property that holds of a theory only relative to some test or tests.

Popper did have other sorts of complaints about psychoanalytic theory. For example, he thought that it was unclear just what its consequences were, and that the consequences it did have were vague. However, though these features adversely affect the testability of a theory, they should not limit the testability of an appropriately precisified version of the theory's claims. And, in any case, such vagueness could be present with a theory having any content at all, and thus is not distinctive of psychoanalysis and Popper's other special targets. It may be, as Popper claimed of psychoanalysis, "practically impossible to describe any human behavior that might not be claimed to be a verification of [this theory]" (Popper 1989, 36). However, the truth of this claim would not be impressive, since any physical behavior might also be adduced to verify properly scientific physical theories if we were prepared to postulate initial conditions and auxiliary hypotheses at will to suit. Serious study of Freud's writings by a reader as critical as Grünbaum led the latter to conclude that Freud was not any more guilty of willful rigging than the typical physicist is. In order to show that there is a special problem with testing psychoanalysis, Popper would need to show that there is a way other than willful rigging of auxiliary assumptions by which psychoanalysis can be made to fit any data. However, the only distinctive feature I can see in this or any other of Popper's examples of problem theories is the self-protective connection between theory and test-auxiliaries on which I have focused.[5]

I am defining lack of candor for a theory in such a way as to allow for the possibility that a theory may lack candor in some tests and not in others, thus making this a local property. In such a case the theory is probabilistically relevant to the probing procedures used in some tests and not to those in others. Such mixed cases—like psychoanalysis on my cursory account—are the most interesting because these theories may be empirically testable despite having a local tendency to self-protection due to the existence of a test-type for which the theory is not candid. Any empirically testable theory could be a part of legitimate science. So there may be theories among those we deem good science whose candor problems, and their consequences, have not been appreciated, and there may be theories among those we regard as evasive of criticism whose testability has been underappreciated due to excessive attention to the tests in which they are not candid.

Whether a theory is candid or not is a question that can be answered only once a test procedure is specified.[6] However, though a theory is testable in principle if there exists a test of it in which the theory is candid, and though this notion of testability in principle might suggest that we take an egalitarian attitude toward the possible tests, commitments and circumstances antecedent to or independent of the theory may give a certain kind of test for a domain a privileged status at a given time (or indefinitely). For example, first person testimony tends to be accorded a privileged status for adjudicating claims about a person's psyche. Naked-eye observations of the heavens had a privileged status in astronomy at the time of Copernicus since they were the only observations of the heavens with demonstrated reliability. Galileo's argument about telescopes came later and was hard-won.

The privileged status of a certain type of test in a domain may hide the testability of a theory from us if the theory is not candid for that test. Thus, ironically, one of the arguments Copernicus developed to show that his proposed new theory of the heavens should not be regarded as absurd also suggested that the new theory might be untestable by the standards of the day. Copernicus noticed that if you supposed that for some reason we would not feel the motion of a moving earth, then naked-eye observations of the daily rotation of the heavens would not be decisive against his idea that the earth moved. That is, the appearances in the sky would be the same, roughly speaking, with heavenly bodies appearing to rotate, whether it was the earth or the heavens rotating (Copernicus 1952, I, 5). This implies that the Copernican hypothesis of heliocentrism, understood as including a supposition that we would not feel the earth's motion if it did move, had not yet been tested by the observations that had been made, and this was Copernicus's point.

However, the very same fact that suggests that the hypothesis had not been properly ruled out, also suggests that it *could* not be tested at the time it was being proposed, because the right kinds of tests to distinguish the two hypotheses had not been described or even considered, and even if they had, the naked-eye observation level of accuracy might not have been sufficient to carry out those tests decisively. It was also the case that the Copernican hypothesis was not candid for the test that he claimed failed to rule it out. The hypothesis implied something about the process of producing the data (observations) in the test that was crucial to the conclusion that the test did not rule the hypothesis out: it implied that even if the earth moved we would not feel its motion. Since it was evident that we do not in fact feel the earth moving, the observations available at the time would, of course, have ruled out the heliocentric hypothesis had it not been for this saving feature of the new idea, a feature which was not justified until long after Copernicus. A local candor problem can take on a global significance in certain circumstances, or at a given time: if the test in which the theory is not candid is the only available or acceptable test there or then, what is strictly speaking a local problem might as well be global.

If a theory lacks candor, and this is a problem (a condition I have not yet explored), then the problem will be of such a character that we can imagine the following consequence: advocates and opponents of a theory to whom only noncandid ways of testing that theory are available may have a hard time settling their disputes, and may each feel the other side is begging the question. This is the kind of phenomenon Kuhn describes as happening when a scientific revolution is afoot. The root cause of the phenomenon I have described would not be psychological dispositions to bullheadedness, having different meanings in mind for key terms, or even the psychological property that human minds are not configured for holding two paradigms at once for comparison. Those are diagnoses that have so far been offered for revolutions. Rather the cause would be a property of the particular theory in the test situations available at a given time: probabilistic relevance relations between the theory and test auxiliaries.

Lack of candor in particular tests is a notion that gives an alternative gloss to some aspects of the notion of paradigm. When it seems as if a theory and its confirming evidence must be accepted all at once or not at all if a science is to proceed at a given point in history, it may be because the theory is relevant to the processes occurring in the test procedures by which it can be imagined to be tested. Such a situation would imply that the probability a person assigns to a theory has consequences for the probability that must be assigned to certain auxiliaries, consequences that are unavoidable if the person is to be rational. This gives a new meaning to the idea that a theory and the methods of probing to discover whether it is adequate are available only as a package deal, and suggests that what it is for a theory to be a paradigm is not always merely that we have an unflinching allegiance to it. Thus, candor problems of a theory might provide explanations different from Kuhn's for cases in which changing one's allegiance from one theory to another looks, at a given point in history, as if it must take the form of a conversion experience.

This suggests a possibility that will be congenial to anyone who fears that acknowledging the existence of revolutions in science will lead inescapably to relativism: it may be that revolutions, with insufficiently grounded changes of allegiance, do happen in science, but because a theory that lacks candor in some tests need not lack candor in all tests, research may genuinely vindicate a theory chosen by revolution through tests imagined only subsequent to the revolution. It is often assumed that if a change of allegiance to a new paradigm is not fully legitimated at the time at which the change of allegiance happens, then it can never be legitimated, because once allegiances change all evidence is seen only from the new paradigm's point of view. Because lack of candor can be local while seeming for a time as if it were global it is possible to imagine eventually overcoming the problems that lack of candor might create: often, later, new tests can be devised. Another feature of candor and its lack, as I have defined them, is important in underwriting this possibility: neither of these properties depends on our allegiances. Because the problems that lack of candor creates do not come from our "seeing things in a certain way" or our prejudices, but from objective aspects of the theories and tests, it is possible to imagine solutions to these problems that do not depend on our "seeing" things in a new way, or on changing or splitting our allegiances either. I will not develop these suggestions further here, but will focus on questions about testability, and about the conditions under which lack of candor is an epistemic problem.

2. WHEN IS LACK OF CANDOR A PROBLEM?

Imagine a situation where we are testing a theory by operating an instrument in observing a phenomenon, or using a probe to measure a phenomenon, or using a device to detect a phenomenon. For example, in an accelerator experiment, we could be using a detector to observe events that will indicate the existence of a certain particle if that particle exists. The theory that we are testing will, in conjunction with auxiliary assumptions, make predictions about what the outcome of this experiment will be if the theory is true. For example, it will predict that the experiment will produce particle events of a specified sort with a specified frequency. If the observational equipment shows an outcome that matches what the theory predicted, let us call such an outcome 'favorable,' and, if the outcome does not match, 'unfavorable.'[7] We will have an epistemic interest not only in whether the outcome of the test was prima facie favorable or unfavorable to the theory, but also in whether the outcome was reliable or not. That is, we will be interested in whether the outcome indicates what we expect outcomes of that sort to indicate, which usually depends centrally on whether the observational equipment functions as expected. Thus, for a given outcome there are four possible ordered pairs of the foregoing two-valued features:

Normally, we should discount pairings 3 and 4, since they were unreliable, and focus on pairings 1 and 2. Outcomes of type 1 will tend to confirm the theory, while those of type 2 will tend to falsify it.

TABLE I

Pairings of outcome properties

Pairing	1	2	3	4
Outcome favorable	yes	no	yes	no
Outcome reliable	yes	yes	no	no

What should we do in the special case where the theory in question lacks candor in this test? Its lacking candor implies that there is probabilistic relevance between the theory and an auxiliary about an instrument, process, or method that produces evidence used in the testing of the theory. These auxiliaries will often include statements that were used in drawing out predictions from the theory; some of them describe the test situation, for without knowing certain things about the testing situation we could not know what the theory implies about it. The auxiliaries would therefore have been used to determine what sort of outcome is prima facie favorable to the theory. They would help us decide things such as 'In this situation, the theory predicts ten blue events per hour and no red ones.'

But many of these auxiliaries will have another function: helping us to determine which outcomes are reliable ('reliability auxiliaries'). Many statements about the process (in the experiment) that produced an outcome will be relevant to whether that outcome was reliable. For example, most statements about how the particle detector works, and worked on the occasion of a particular test, will be relevant to whether observed events or absence of observed events means that there were or were not particles. In the most obvious kind of case, if the particle detector was not working as planned, then an absence of detected events is not an indicator that the sought-for particles do not exist.

In a normal testing situation, lots of statements about the process that produces the outcome will be probabilistically relevant to the reliability of the outcome. For example, optical theory will be relevant to the process that goes on in a light microscope, and will help us to determine of a given appearance whether it is an artifact or not. However, light microscopes are often and fruitfully used to test claims about things other than optical theory. For example, we can use them to study living cells and organisms. In those cases, although there is a theory that is relevant to the reliability of the test process, i.e., optical theory, this theory is not the theory under test, and so no lack of candor arises.

If the theory under test fulfills the first condition above for lacking candor, then that very theory under test is probabilistically relevant to some of the auxiliaries about the process that produced the outcome. This sounds fishy, but in analyzing its significance, one must proceed carefully. For example, we might expect that this situation makes the theory relevant to the reliability of the outcome, but it does not automatically do so. Probabilistic relevance is not in general transitive.[8] So probabilistic relevance of a theory to the process used to produce outcomes can fail to produce a problem for testability in the following way: though the theory is relevant to the process that produced the outcome, and the process that produced the outcome is relevant to the reliability of the outcome, the theory is not relevant to the reliability of the outcome.

However, in many cases the transitivity condition will be fulfilled. In those cases, the theory is probabilistically relevant to the reliability of the outcome. I will take a theory to *lack candor* in a test situation if the theory is (1) relevant to an auxiliary about the test process in a test situation, and (2) (thereby) relevant to the reliability of the outcome. With individually necessary and jointly sufficient conditions in hand for when a theory lacks candor, our question now is whether fulfillment of these conditions by a theory and a given test situation means that we should respond differently from normal to the outcomes of the test.[9] The view I will defend here is that our default attitude should be to regard tests in which theories lack candor with suspicion—in this sense, Popper was right—but that it is also often possible for us to find good reason to set that suspicion aside. I will discuss my reasons for the first claim in this section, and those for the second in §3.

One might think that it is just obvious that we should regard such tests with suspicion. If so, it is probably because of the thought that the intuitive relevance of the theory to the reliability of the outcomes means that we will be forced to reason circularly when we take the outcomes as indicative of whether or not the theory is true. However, that the theory is intuitively relevant to the reliability of the outcomes does not imply that we will reason circularly in taking an outcome as either reliable or unreliable. It is logically possible that we believe the theory before accepting an outcome as reliable, and that that belief is relevant—in an intuitive sense—to the reliability of the outcome, yet our belief in the theory is not the *basis* for our belief that the outcome was reliable. We may have another, independent, belief that is the basis.

This way of avoiding circularity even when there is some type of relevance between reliability of outcome and truth of the hypothesis under test does not at first sight appear to be available when we think about the matter in terms of probabilistic relevance. Few doubt that positive probabilistic relevance of e to H is a necessary condition for e to be evidence (support) for H unless they reject altogether the project of representing evidence via probability (though see Achinstein 2001; Roush 2004a). However, if we add, as many would, that e's being probabilistically relevant to H is sufficient for e to support H, then the threat of circularity in cases where a hypothesis lacks candor seems acute. For if H is probabilistically relevant to the reliability of e, as we said it was in any case where H lacks candor, then the only thing that prevents H from illegitimately boosting itself is failures of transitivity in which H is relevant to e via the auxiliary, and e is relevant to H (directly) but H does not thereby get to boost its own probability.[10] Failure of transitivity in such situations can happen, but it hardly seems common enough to engender confidence that we will always avoid circular reasoning with noncandid theories.

One might respond to this conundrum in three different ways. First, one might deny that noncandid theories should be counted as evidence for themselves and reject the idea that positive probabilistic relevance is sufficient for evidence. Second, one might embrace the conclusion that noncandid theories should count as evidence for themselves and maintain the view that positive (probabilistic) relevance is sufficient for one claim to be evidence for another. Neither of these proposals seems especially attractive. The third broad type of response is to hope that there has been some misunderstanding in the derivation of this dilemma. Indeed, as I will argue in §3, we are led into this dilemma only if we fail to appreciate some important details about how probabilistic relevance works, and how it is different from intuitive relevance. Thus, as I will argue, there are factors in addition to failure of transitivity that explain why a theory's positive probabilistic relevance to an outcome's reliability (when the two are considered in isolation) does not automatically make that theory count as evidence that the outcome is reliable.

Some will not be convinced that this third option is needed, because they are not convinced that lack of candor, or even circularity, is a problem. In order to explain why the second option, in which we maintain the positive relevance conception of evidence and stop worrying about lack of candor, is unacceptable, I need to move back a step and argue, in the remainder of this section, that our default attitude toward noncandid theories should be suspicion. In recent years some authors have strenuously argued that the widespread presumption against tests in which there is relevance between a theory and an auxiliary used to test the theory, and even our squeamishness about circularity, are prejudices that should be given up. On this view as described by Harold Brown, we should consider each case individually in detail before condemning any of this kind (Brown 1994, 408, 410). Hudson (1994, 605; 2000, 9–10) even thinks that a scientist's willingness to question the theory under test *if untoward results arise* will insure that what he calls "background dependence" in a test of the theory is wholly unproblematic:

> [D]espite the fact that one's observations are informed by a theory T, so long as an experimentalist is willing to question T, given untoward results, there is no need to be concerned about vicious circularity in one's testing procedure. The crux is the experimenter's *attitude* toward the testing situation, to wit, is she genuinely open to questioning the theory informing the observations? (Hudson 2000, 10)

Hudson thinks that the right attitude makes even circularity innocuous, so it seems that he would certainly think that the relevance relation I have indicated—which may or may not yield circularity—is unproblematic. I disagree with both of these authors. I will argue that in a situation where there is probabilistic relevance between a theory and auxiliaries used to test the theory, and nothing else is known, the relevance makes the test unrevealing (to an extent related to the degree of relevance) in every possible case. Thereby I will defend my view that our default attitude toward situations involving such relevance should be suspicion.

Hudson may be thinking that we can trust untoward results if the theory has been assumed in producing them because if the theory was assumed, the theory could only have been working in its own favor. If disconfirming results came through despite this force trying to hide them, then they must be indicative of the nature of things, or at least not due to the theory's own action. But even if this rationale is sensible, it does nothing to reassure us that we can trust *favorable* results in which the theory has been assumed. Indeed, if the theory is assumed and is thereby working in its own favor, then to that extent it is not possible to tell whether a favorable result is indicative of the way the world is or is a consequence of the theory's own prejudicial work. The experimenter's attitude can do nothing to console us about this, since it is a matter of the test not giving us information.

Imagine the theory that Acme makes reliable, accurate clocks, and that you test this theory by sitting in a windowless room with one of their clocks, watching it. Suppose that there is no other timepiece in the room. You sit and observe this clock for hours, thus losing any independent sense of what time it is, and all this time the clock continues to run apparently uniformly without stopping.

Obviously, although you have been given no indication that the clock is inaccurate, you also have no evidence that the clock is accurate since you have no independent check on the time. You could assume as an auxiliary in your test that Acme makes reliable, accurate clocks, and thus that since the clock had the right time when you went into the room it probably has the right time now too. However, this would be manifestly circular if the point of the exercise were to *test* whether Acme produces clocks that maintain correct time. Notice too that having a good attitude does nothing to change this situation: a willingness to question the theory if untoward results about accuracy of the clock arise is of no consequence, since untoward results will not arise. The fact that untoward results never actually arise in this test does not yield confirmation of the theory, and an inquirer's questioning outlook makes no difference to this.

There is a further problem with Hudson's assurances, for neither untoward results nor toward results can be taken simply as "given"—as if they are missives straight from the nature of things—when relevance of the sort I am considering obtains, because whether the results are reliable, and therefore whether they should count, depends on matters to which the theory is relevant. With this kind of relevance we will find that in the default situation, to the extent that there is relevance between theory and reliability auxiliary, the possible test outcomes are unrevealing. This is so despite the fact that the work that the theory does in the test is not in every case in its own favor.

The simplest, default situation gives us an argument that our default attitude toward tests lacking candor should be suspicion. In the default situation we know what the theory is, we know what test we are doing of that theory and what auxiliary assumptions are crucial to that test, and we know the test outcomes. Crucially, we do not know anything else that is probabilistically relevant to the outcomes, or to the auxiliaries to which the theory is relevant, or to the reliability of the outcomes.[11] From this it seems to follow, in particular, that we do not have independent evidence for the auxiliaries to which the theory is relevant or for the reliability of the outcomes, regardless of how we might make more precise the intuitive idea of independent evidence.[12] The only reasons that we will have to believe in the truth of some rather than other auxiliaries relevant to reliability of the test outcomes will be reasons that include beliefs about whether the theory being tested is true; even if the probability of the theory does not fully determine the probability of the auxiliary, it is the only thing that can affect the latter. Beliefs about whether the theory is true will be our basis for believing auxiliaries relevant to the reliability of outcomes if we have any basis at all. In this case our epistemic reasons will follow the probabilistic relevance relations between the theory and auxiliaries relevant to the reliability of outcomes rather closely. Any increase or decrease in our estimation of the probability of the theory will translate immediately into an increase or decrease (not necessarily, respectively) in our estimation of the probability of the reliability of the outcomes. This situation does yield intuitive epistemic circles, and their details are worth exploring.

The fact that theory and reliability of the test outcome are probabilistically relevant to each other means that the probability of the reliability of the outcome is sensitive to changes in the probability of the theory.[13] The theory may be either positively or negatively relevant to the reliability of outcomes. Consider the positive case, where two intuitive circles will emerge. Suppose we have a favorable outcome. Is it reliable? The theory counsels us that it is, since the theory is positively relevant to the claim that this outcome is reliable and we have no other information. (That is, the prior probability of the auxiliary is .5.) A favorable outcome that is judged reliable is more likely to confirm the theory (because more likely to count), than if the outcome had been judged unreliable. Thus, it looks as if in such a case the theory gets to boost its own probability by reinforcing outcomes confirmatory to itself. Importantly, this is not actually what happens probabilistically. Conformably to our sense that this is a kind of circular boosting that should not be allowed, the equations of probability keep track of and wash out double-counting, and the actual upshot in such a case would be no confirmation or disconfirmation of the hypothesis from this evidence at all. This is because the only information that licenses counting the evidence is information that was already in the hypothesis. Probability conforms to our intuitions here about circular reasoning.

When we get an unfavorable outcome, the theory that is positively relevant to the auxiliary that says this outcome is reliable counsels us to count this outcome at face value. If an unfavorable outcome is reliable this decreases the probability of the theory a little because it is a disconfirming instance. If, as we are assuming, the theory is positively relevant to the reliability of the outcome, then if the probability of the theory decreases, the probability of reliability of the unfavorable outcome decreases too. If the probability of the reliability of the outcome decreases then the disconfirming force of the outcome on the theory decreases, raising the probability of the theory. It appears, then, that the theory will have a see-sawing cyclical effect on its own probability. Of course, this does not actually happen if we follow the probability calculus since, once again, the equations of probability keep track of double-counting, and no effect on the reliability of the evidence that came only from the theory would yield a confirmation or disconfirmation of the theory (change in probability of the theory) by that evidence.

There are also two circles in case the theory is negatively relevant to the reliability of the outcome. Suppose a favorable outcome. The theory tends to decrease the probability of the reliability of the outcome. The more unreliable a favorable outcome is deemed to be, the less chance that favorable outcome has to confirm the theory compared to what it would have done had it been deemed reliable. If the theory is boosted at all then so is the claim that the favorable outcome is unreliable. If we kept going in this circle, it seems, all confirming power of the favorable outcome would be destroyed. In fact, the outcome can have no confirming or disconfirming effect on the theory, but this obtains for a different reason: if the only information that tells us whether or not to count the outcome comes from the theory, then it would be double-counting to let that affect whether the outcome gets to boost the theory, and the equations of probability automatically ensure that this does not happen.

In case the theory is negatively relevant to the reliability of an unfavorable outcome, the theory will counsel against the reliability of this outcome. The less likely that an unfavorable outcome is reliable, the less it will be counted as disconfirming a theory. In this case to follow the theory's advice would be to shield the theory from disconfirmation on no recommendation other than that of the theory itself. Waxing anthropomorphic, we might say that it would be to allow the theory to engage in self-protection. The equations of probability conform to our sense that this is not kosher, for if the only information tending to show that a piece of evidence does not disconfirm a theory comes from the theory itself then it would be double-counting the theory to register this positively for the theory, and using the probability equations properly takes care of this without our thinking about it.[14]

The first thing to notice from all of this is that a theory's relevance to the process producing the outcome would not always serve the theory if we were allow the theory to affect its own confirmation. It is not the case that relevance always yields a propensity to self-protection. The second thing to notice is that the effect a theory would have on its own fate is bad in every case from the point of view of someone trying to use this test to judge whether the theory is true. We cannot trust the favorable outcome in the case of negative relevance between theory and reliability of outcome, because the theory itself tells us not to do so. That is, even if we gave the theory the benefit of the doubt by assuming it, we would find that this outcome favorable to the theory should not be trusted, so the outcome should not be trusted. Because of epistemic circularity, we cannot trust a favorable outcome in a case of positive relevance between theory and reliability of outcome: the theory tells us that we should trust the favorable outcome, and it is our only relevant source of reasons, so the theory would be our basis for taking something to be evidence for it. When the theory tells us not to credit an outcome unfavorable to it, we similarly cannot employ its advice without begging the question. On the other hand, in the default situation we are considering, we have no other way of judging the reliability of that unfavorable outcome, so we cannot reasonably credit it either. This leaves us in a situation where the test tells us nothing. When the theory is positively relevant to the reliability of an unfavorable outcome the effect of the theory on its own confirmation, if we allowed such an effect, would be oscillating, therefore undefined, and of no use, as we saw above. All of these epistemic judgments have counterparts in the fact that probability theory automatically washes out double-counting and would not allow a theory to boost or depress its own probability by boosting or depressing an intermediate claim. Interestingly, though self-protectiveness of theories has been noticed more often in examples, self-sabotage also interferes with testing a theory.[15]

I do not claim that the stripped down scenario that I have called the 'default situation,' in which no information other than the relevant theory is available, is common or uncommon in actual science, though it seems to be possible. However, if other information than what I allowed were available, the relevance relations I have considered would still obtain if the theory and the auxiliary were considered in isolation. If we did not know the import of the other available information we would not know whether it overcame the problems that I have associated with the relevance in question. Thus what I have just argued shows that if we discover probabilistic relevance between a theory and the reliability of outcomes of the procedures for testing it, then we should regard those tests with suspicion until it is shown how other available evidence overcomes the problem. This means that, *contra* Brown, we should not have a neutral attitude toward theories that lack candor until we see the details of cases; we should regard theories that lack candor as guilty until shown to be innocent.

3. CAN THE PROBLEMS CAUSED BY LACK OF CANDOR BE OVERCOME?

In §2, in discussing tests in which the theory under test is relevant to an auxiliary about an instrument or method involved in producing the outcome of the test (a reliability auxiliary), and thereby relevant to the reliability of the test's outcome, I assumed we did not have evidence for the auxiliary that was independent of belief in the theory. If we did have independent evidence for the auxiliary, then the trouble that the probabilistic dependence between theory and auxiliary causes might be tempered, since the probability assigned to the auxiliary would not depend exclusively on the probability assigned to the theory, but would also depend on that other evidence. What must "independent evidence" mean for this to be so? Here our intuitions may balk. To say that a theory and reliability auxiliary are probabilistically relevant to each other (for probabilistic relevance is symmetric, so the theory's relevance to the auxiliary is reciprocated) is to say that the two are not probabilistically independent. How then could we have evidence for one that was independent of the other if being evidence is a matter of probabilistic relevance, i.e., lack of probabilistic independence?

The first thing to say is that even if being evidence for X is the same as being (positively) probabilistically relevant to X, that does not imply that being independent evidence for X implies being probabilistically independent of something. There are notions of independence concerning evidence, definable in probabilistic terms, that are not the same as probabilistic independence simpliciter.

There are at least two plausible notions of independent evidence for the reliability auxiliary that would allow there to be a kind of independent evidence even where there was not probabilistic independence between the theory and the auxiliary.[16] The first is the notion of a claim being evidence for the auxiliary independently of the theory. That is, first, the evidence claim e meets a necessary condition for being evidence for the auxiliary A, and second, the effect of the evidence on the auxiliary is independent of the theory T. These conditions can be represented as follows:

(1) $\mathrm{P}(A/e) > \mathrm{P}(A)$,

(2a) $\mathrm{P}(A/e.T) = \mathrm{P}(A/e)$,

(2b) $\mathrm{P}(e/A.T) = \mathrm{P}(e/A)$.

The first condition, which says that e is positively relevant to A, thereby says that e meets the popular requirement discussed earlier for being evidence for A. Conditions (2a) and (2b) are alternative (and inequivalent) ways of demanding that e's meeting this requirement for being evidence for A does not depend on T.[17] (2a) says that T makes no difference to whether e raises A's probability. (2b) says that T makes no difference to whether A raises e's probability.

The probabilistic relations required by (2a) and (2b) are called 'screening off' relations. Z *screens off* Y from X if and only if $\mathrm{P}(X/Z) = \mathrm{P}(X/Z.Y) \neq \mathrm{P}(X/Y)$. In such situations, though Y may be probabilistically relevant to X when the two are considered in isolation, and though Z does not change that fact, when Z screens off Y from X, assuming Z true renders that relation between X and Y, as it were, impotent.[18] It is sometimes said that Z "shields" X from Y.[19] Moreover, if Z screens off Y from X, then according to a probability function that assigns probability 1 to Z (effectively taking Z to be true), the relevance between Y and X is not merely shielded but nonexistent, since Y and X are *not probabilistically relevant*. In our case, the relation between T and A remains when the two are considered in isolation, but given the type of independent evidence we are discussing, either e screens off T from A, as in (2a), or A screens off T from e, as in (2b). Moreover, for a function P' for which $\mathrm{P}'(e) = 1$, $\mathrm{P}'(A/T) = \mathrm{P}'A)$, implying that with this function T is not probabilistically relevant to A (equivalently, A is probabilistically independent of T), and for a function P'' for which $\mathrm{P}''(A) = 1$, $\mathrm{P}''(e/T) = \mathrm{P}''(e)$, implying that with this function T is not probabilistically relevant to e (equivalently, e is probabilistically independent of T).

These conditions for independent evidence can be fulfilled by a theory, reliability auxiliary, and test in which the theory and auxiliary are probabilistically relevant to each other (when considered in isolation). For an example that fulfills the conditions in the limit as evidence accumulates, that is, as we approach examining all the instances relevant to the reliability auxiliary, consider the theory T:

All fluids expand on increase of their temperature.

Consider the test of T in which as we heat many fluids, we measure their volumes and take their temperatures with a mercury thermometer. One auxiliary hypothesis, A, that we want to assume about the instrument is obvious:

Mercury expands on increase of its temperature.

But T and reliability auxiliary A are obviously relevant to each other (considered in isolation) since the latter is a simple restriction of the former, assuming as we can from background knowledge that mercury is a fluid. Despite this, we can have evidence for the auxiliary that is independent of the theory (in the limit) in the sense just defined. Intuitively, this is because there are other ways of measuring temperature than glass bulb thermometers that rely on volume.[20] We need to see how the notion of probabilistic relevance respects this idea.

Consider as e the claim that the electrical resistance method of measuring says that the temperature of the thermometer is, say, 72° now and was 62° five minutes ago (when the volume was seen to be smaller). Suppose these values match what the thermometer registered as its temperature at these times. Clearly this e is probabilistically relevant to A, the claim that mercury expands on heating, at least on the assumption that the electrical resistance method is not so bad a method of measuring temperature that it gives random results, and the assumption that T has not been assigned a probability of 1 (which would make it impossible for anything to be evidence for it). This shows that condition (1) is fulfilled. A moment's reflection also shows that condition (2b) is fulfilled in the same circumstances. The relevance of A to e is unaffected by whether T is true (i.e., A screens off T from e), because the only samples e refers to are samples of mercury, and A says about mercury everything that T says about mercury. T is a wheel that doesn't turn anything in the relation between e and A.

Fulfillment of condition (2a) is a more complicated issue. The probability of the auxiliary A given the evidence e is to some extent unaffected by whether the theory T is true, assuming the electrical resistance method does not have random outputs, because e gives some of the boost to the auxiliary that it could otherwise have gotten from the theory. However, since in this case the theory is not merely relevant to the auxiliary but *implies* it, the evidence e would have to raise the probability of the auxiliary to 1 in order fully to replace the effect of the theory and fulfill condition (2a). This is the most extreme kind of relevance there could be between T and A, and where the relevance was less extreme, condition (2a) would be easier to fulfill.[21] But even in this extreme case we can see that in the limit where we witness, and therefore insert as 'e' all instances relevant to A, we will approach fulfillment of condition (2a). Even when we are not at the limit, we can see in what way the evidence from the alternate method of measuring temperature is acting: so as to replace step by step the effect of prior assumptions about the theory on the probability of A.

Because it is possible for evidence to screen off the relevance between a theory and an auxiliary, it is possible for there to be evidence that supports the auxiliary independently of the theory. This is a chief ingredient in overcoming the problem presented by lack of candor in a theory. A theory that is probabilistically relevant to a reliability auxiliary when the two statements are considered in isolation, and which thereby lacks candor for the test in question, may cease to be probabilistically relevant to the auxiliary when more evidence is taken into account. Such a theory will cease to be probabilistically relevant to the auxiliary, according to the probability function that results from taking that evidence into account, if the further evidence is such that it screens off the theory from the auxiliary.

To see this clearly one must consider carefully an important difference in the ways that intuitive relevance and probabilistic relevance get evaluated. When we ask whether Y is relevant to X in the intuitive sense, the default assumption is that we are considering the two statements in isolation. The assumption of isolation can be withdrawn, but this must be done explicitly by addition of a phrase like 'given that Z.' In this intuitive sense of relevance, T does not cease to be relevant to A when e screens off T from A, for example, because nothing has changed about the fact that T and A have overlapping subject matters, and the default procedure is to consider the two statements in isolation. It is quite otherwise with probabilistic relevance, where pains would have to be taken to express explicitly what considering two statements in isolation means. Since the condition for probabilistic relevance between Z and X is that $P(X/Z) \neq$

$P(X)$, whether two statements are probabilistically relevant can only be answered once a probability function is specified. A probability function assigns probabilities to every statement in the language. The question, then, of whether Z is probabilistically relevant to X is the question whether the truth of Z would affect the probability of X when the given assignments of probabilities to all other statements is taken into account. Thus, if we say that e screens off T from A, we are saying that in a probability function that assigns probability 1 to e, and probability 1 to the statement that the electrical resistance method does not produce random outputs, T is not probabilistically relevant to A. That is, if we know e (and something about the electrical resistance method), T no longer fulfills the conditions for a theory to lack candor. (Note that on this particular point the analogy between people and theories with respect to their candor breaks down, for even if we find evidence independent of a person's noncandid testimony which helps us to settle the matter testified to, that will not make the person seem any more candid.)

The statement of this conclusion brings out the fact that the kind of independence of evidence I have just shown can be secured for an auxiliary even when testing a theory that lacks candor is not all that we could hope for. The conclusion is a conditional statement, and T ceases to lack candor only if we *know* e and something about electrical resistance. Can we know e independently of T? Put differently, while the screening off relations (1), (2a), and (2b) assure us that T does not affect the relevance between A and e, e will not actually boost A much unless e has a relatively high probability. Where does the justification for that high probability for e come from, and can it be had without making assumptions about T? A can be confirmed independently of T only by something with screening off properties like those of e, but we've just pushed the question back one step because we can now ask how e is to be confirmed, and whether that can be accomplished independently of T.

At first sight the prospects look grim. Notice that the assumption that made e relevant to A, namely, that electrical resistance is better than random at measuring temperature, also made e relevant to T, since that made e verify an instance of T (to some degree). If e is relevant to T when the two are considered in isolation, then T is relevant to e when the two are considered in isolation, and therefore T is a candidate for acting as evidence for e. That is, e is evidence for A independently of T, but is not itself probabilistically independent of T when the two are considered in isolation (i.e., it is not the case that $P(e/T)=P(T)$, if A is not assumed).[22] This is actually required by the definition for screening off to obtain, but does it mean that we cannot know e independently of T?

Intuitively we would probably say "no." After all, the intuition that says that e, the evidence for A, should be "independent" of T is met by our being able to *know* e without knowing T. We can know that the electrical resistance method of measuring says that the temperature of the thermometer is 72° now and was 62° five minutes ago without knowing that all fluids expand on heating. From our ability to know e without knowing T, it does not follow that e is probabilistically independent of T. To take a simpler example, we may know an instance without knowing its generalization, though the two are probabilistically relevant to each other.

Nevertheless, it remains to see whether probabilistic relevance respects our intuitions here, and if so, how. In the case described, we come to be confident of e and of its import for A because of investigations by and of the electrical resistance method of measuring temperature. e itself is actually easy to verify without regard for the truth or falsity of T, because e is only a report of what the electrical resistance method said. That is, we may conditionalize on e (assign it probability 1) because of inspection of an instrument, and the outcome of this inspection will not depend on the probability assigned to T. However, the import of e for A depends on electrical resistance being a *reliable* way of measuring temperature, and here the intuitive worry about T acting as evidence for itself finds its technical counterpart. Call the assumption that electrical resistance gives a reliable way of measuring temperature 'E,' in view of the way this assumption underpins the significance of e for A. E must be known if the condition $P(A/e.T)=P(A/e)$ is to be fulfilled. Can we learn E without any regard for the truth or falsity of T?

Let us imagine that our electrical resistance apparatus does not take volume measurements on the two samples but rather identifies the samples by the time at which they are measured. (The mercury thermometer itself is our source of information about the volume of the mercury.) This is possible because neither the electrical resistance method of measurement of temperature nor the method of calibrating this method relies on information about volume. Temperature gets its qualitative meaning from sensations of touch and other phenomena (boiling, steam) rather than visual observation of volume changes. Establishing the quantitative scale is more delicate, but this is officially accomplished by bringing substances to special qualitative states whose maintenance indicates constancy of temperature, e.g., the triple point of water, and extrapolating a scale between such points. We could calibrate our electrical resistance method with this scale.

If so, then reliability of electrical resistance as a measurer of temperature is probabilistically independent of T, since electrical resistance reports about temperature do not depend on volume or say anything about volume. Both T and E are probabilistically relevant to e and e is relevant to each of T and E when these terms are considered pairwise in isolation. However, even in isolation E and T are not probabilistically relevant to each other. This means that what lies at the basis of our ability to avoid circularity in cases like the one I described is a failure of transitivity of probabilistic relevance. Our confidence in the import of e rests on E, which is probabilistically independent of T, even when E and T are considered in isolation. This aspect of the situation shows that it is a welcome failure of transitivity, and the screening off relations discussed above, that make it possible to get independent evidence for an auxiliary even when the theory under test was not candid for the test when we imagined ourselves in the "default" situation. Because of the existence of such independent evidence the theory ceases to lack candor in the test.

We started off noticing that we could have a belief Y that was intuitively relevant to statement X but not the basis for our belief in X, and doubting that the probabilistic relevance notion of evidence could capture this situation, since to be evidence just *is* to be relevant on that view. However, as I have emphasized, intuitive relevance and probabilistic relevance often differ in cases where the background assumptions are not explicitly identified, because the two modes of evaluation have different defaults. This means that we will not get the right answers about what is evidence for what according to the positive relevance notion of evidence if we suppose that that notion of evidence uses an intuitive concept of relevance. What probabilistic relevance *does* follow closely, and thankfully so, is the intuitive notion of the *basis* of a belief. For example, if we said, as we did early on, that T is intuitively relevant to e but not the basis for belief in e or consequently for belief in A, the reason for that would be that E screened off the relevance between T and e, i.e., because when E is known T is *not probabilistically relevant* to e. Probabilistic relevance closely follows the notion of a basis.

Some readers may be reminded here of work by Salmon (1975) in which he identified the phenomenon responsible for the apparently anomalous cases that arise for a notion of confirmation that uses probabilistic relevance, cases first systematically catalogued by Carnap (1950). In these cases, for example, though a piece of evidence e confirms a hypothesis H and an auxiliary A, it disconfirms their conjunction, or though e disconfirms the disjunction of H and A it confirms each of H and A, and there are many other permutations. Salmon pointed out that in all of the cases the addition of the evidence e changed the level or kind of probabilistic relevance between H and A.

He noted further that it would not make sense to adopt a rule forbidding all situations in which addition of e changed the relevance between H and A, because a standard kind of disconfirmation scenario required such a change. In this scenario, H and A start out probabilistically independent, which Salmon regarded as ideal since we could then have a belief about A without that forcing any particular stand on H. H and A also together imply e, the prediction the hypothesis leads us to expect. However, not-e is what actually occurs. The occurrence of not-e means that H and A are no longer probabilistically independent because since they together imply e they cannot both be true in a situation where e is false; if one of them is true the other must be false, and vice versa. If it can be determined whether H or A is more likely to be the culprit, then a falsification will have occurred. Thus there are scenarios where evidence changes the relevance relation between a hypothesis and an auxiliary, and far from objecting, we should rather applaud this feature.

What we have in the case discussed at length above is another type of example in which we should applaud and not object when the addition of evidence e changes the relevance relation between a hypothesis H and an auxiliary A. In this type of case the addition of evidence e changes the situation from one in which H is probabilistically relevant to a reliability auxiliary A to one in which A is probabilistically independent of H, i.e., neither is probabilistically relevant to the other. e is evidence for the auxiliary A in this case, not principally evidence for H, but the addition of e as evidence makes other evidence for or against H possible because it shields our belief about A from our prejudices about H, and thus allows us to count or discount putative evidence for or against H independently of the influence of our prejudices about H. For this reason it can be a step that precedes the case Salmon focused on, where H and A start out independent.

The counting and discounting of putative evidence, recall, depends unavoidably on judgments about reliability auxiliaries. In Salmon's case the kind of auxiliary in question is what we might call a 'prediction auxiliary' because it is an assumption that must be made in order to draw out predictions from the hypothesis H. The two types of auxiliary assumption play distinct but broadly similar supporting roles in the process of confirmation and disconfirmation: a prediction auxiliary should be independently confirmed to some degree in order for evidence that contradicts the prediction to impugn the hypothesis (and not the auxiliary), and a reliability auxiliary should be independently confirmed to some degree in order for an outcome it says is reliable to be taken to confirm or falsify the hypothesis, and in order for an outcome it says is unreliable to be taken to be of no consequence.

There is another kind of independent evidence that it is possible to have when there is relevance between theory and auxiliary. In this notion we are interested in evidence for the auxiliary that is independent not of the theory but of the *evidence* we have for the theory. This kind of independence promotes variety of evidence. The conditions that must be fulfilled for this are:

(3) $\mathrm{P}(A/e.e') > \mathrm{P}(A/e')$,

(4) $\mathrm{P}(e/A.e') = \mathrm{P}(e/A)$.

Suppose e' is the evidence we already have for our theory and e is the new evidence that we want to boost A independently of the first evidence.[23] For this, we want (3)—that e makes A more probable than it would be without e even when e' is assumed—which ensures that e is evidence for A that is not redundant with e'. And we want (4)—that whether A makes the evidence e more probable does not depend on the old evidence.[24] Condition (4) could also be described as saying that e and e' are independent of each other *relative* to A, a notion discussed by Sober (1989).[25]

We might worry that when, as we now assume, T is relevant to A, these conditions cannot be fulfilled, for since e' is evidence for T it must be relevant to T, so how could it fail to be relevant to e, which, being evidence for A, is relevant to A? But note that in this worry we have assumed twice over that relevance is transitive, which it is not. If the relevance fails the transitivity condition in either case, then e' will be (unconditionally) independent of e. Of course, though it is interesting that this is possible, this is more than we need to fulfill the conditions above which involve relative rather than unconditional independence. These conditions are fulfilled by our thermometer example above.

To see this let e', our old evidence for the temperature-volume theory, be evidence we got, say, from volume measurements of hydrogen taken by us at a variety of temperatures verified by appeal to the black body method. Let e be the same evidence as above that confirms the working of the mercury thermometer (i.e., confirms A) by measuring the thermometer's temperature at various volumes by, let us say, the electrical resistance method. (Suppose it is also found that the temperatures that the thermometer reports match these electrical resistance readings of its temperatures.)

Condition (3) is fulfilled because even if we are given that e'—temperature increases with volume for hydrogen—assuming that this mercury expands on an increase in its temperature according to the better than random electrical resistance method will make it more likely that our mercury thermometer is working as we think it is than that would be if we did not make the assumption. Condition (4) is fulfilled because when the auxiliary proposition that mercury expands on heating is assumed, the fact that hydrogen expands on heating does not make a difference to the probability that temperature has increased with volume, in line with the markings on our mercury thermometer, whatever difference the latter might have made without that assumption.[26]

We could describe these ways of getting independent evidence for the auxiliary about the thermometer as engaging in a different test of the theory than the test with which we began. So instead of testing by measuring volumes and taking their temperatures via the mercury thermometer, we are doing a less informative test by restricting our focus to mercury and using some other method to decide whether indeed this mercury expands on heating. But we can also describe this as the test with which we began, in the sense of attempting to test the same claim, the theory about all fluids, with independent evidence for the auxiliary added to the test. Of course, technically the addition makes it a different test than the one with which we began; addition of the electrical resistance evidence about the thermometer implies fulfillment of the other set of conditions for independent evidence (1), (2a), and (2b) as well. We saw that fulfilling those conditions made a formerly noncandid theory candid, and candor was defined relative to a test. However, this distinction is minor compared to the main points: first, it is possible for a theory that is not candid for a test to become candid in an extension of that test by the addition of independent evidence. Second, there is a further kind of independent evidence, associated with variety of evidence, that it is possible to find for a theory that lacks candor.

To return to the overarching questions about lack of candor, it is not only that there can be other tests one might perform on a theory that is not candid for the test with which one starts, as I suggested above is the case for psychoanalysis. It is also possible in principle to find evidence that overcomes the problem that lack of candor creates in the test with which one began. This makes it a different test, technically, but one is still following the same route of testing. Notice also that this has been shown without giving up a positive relevance notion of evidence. Since lack of candor is a matter of relevance, upholding a notion of evidence to which relevance is central has given the widest possible berth to the idea that lack of candor could create circularity in evidence: the idea that a noncandid theory will contribute to the evidence for that very theory. The fact that the results of this paper leave in place the positive relevance notion of evidence increases the force of the conclusion that though lack of candor in a theory will make testing it a challenge, it does not necessarily make this impossible.

4. CONCLUSION: BACKGROUND INDEPENDENCE

A number of people have thought about issues related to those that I have been discussing. Sometimes they have called the demand referred to earlier that "in testing a theory, one should use observations whose theoretical underpinning excludes the theory under test" the requirement of *background independence* (Hudson 1994, 595). Two key problems have hobbled some of this discussion. One is the ambiguity of phrases like 'theoretical underpinning excludes...' said of observations used as evidence, and its opposites in which observations are 'informed by theories' and 'presuppose theories.' The second difficulty is the mistaken assumption that there is a single question posed using these phrases.

One axis of ambiguity in these phrases is the distinction between epistemic independence and factual independence. Observations that are 'informed by theories' pretty clearly seem to fail to be *epistemically* independent of those theories, while observations whose 'theoretical underpinning excludes' a theory seem to achieve this independence from the theory.[27] This is because these phrases connote our grounds for believing the observation statements in question. However, the phrase 'x presupposes y' has at least two common meanings, one of which connotes not relations between beliefs and reasons for beliefs, but relations between matters of fact: x *presupposes* y just in case in order for x to be true, y must be true, i.e., just in case x entails y. (A probabilistic version of this, which would have to assume an objective interpretation of probability, would say that x makes y probable, or x increases y's probability, i.e., x is positively probabilistically relevant to y.) One epistemic reading of 'x presupposes y,' the one that Sober had in mind when he claimed that "the theories under test cannot be presupposed by the observation statements that are used to test those very theories," has it mean something such as 'you can't know that x is true unless you already believe y' (Sober 1999, 52).[28] (A probabilistic version of this would say that you probably cannot know that X is true without knowing that Y is true.)

It seems to me that the guiding question about independence of observations or experimental outcomes from the theories that they are used to test is epistemic. Tests are used to produce grounds for accepting or rejecting theories, and what we want to know is whether dependences, of whatever sort, interfere with that epistemic task.[29] Dependence between matters of fact has entered the discussion of these topics so far mainly unconsciously, through not paying sufficient attention to the difference between relations among matters of fact and relations between reasons and beliefs, or because epistemic dependence follows probabilistic relevance closely when the latter is properly understood, and probability can be interpreted objectively. In this paper I have distinguished the intuitive notions of basis of a belief and relevance between two statements from their probabilistic counterparts, and investigated how these two registers are related to each other. Only so were we able to understand how it is possible to overcome the epistemic challenge that a theory's lack of candor creates.

In particular, we saw that observations may be factually dependent on the theory under test, in the sense that the theory is probabilistically relevant to a reliability auxiliary when those two statements are considered in isolation, yet not epistemically dependent on the theory under test if there is independent evidence for the auxiliary. This is because the independent evidence screens off the relevance between theory and auxiliary, meaning that once that evidence is taken into account the theory no longer supports the auxiliary epistemically. Although the theory and auxiliary are not probabilistically relevant to each other once that evidence is taken into account, they remain probabilistically relevant to each other when considered in isolation (factually dependent). Thus, once the independent evidence is in, the auxiliary is epistemically independent of the theory, but this does not change the factual dependence between the two. If we fail to appreciate these distinctions we cannot avoid confusion in discussing so-called 'background dependence.'

We can now answer some questions about whether, under each of the plausible univocal meanings of the phrase, background independence should be required in testing. Start on the conservative side of the spectrum of views, and consider Sober's criterion, under the first reading of 'presupposes' (which is not the reading that he intended). This says that a theory must not be presupposed by observation statements used to test it, and means that the theory and, I will suppose, its logical consequences, must not be entailed by those observation statements. Must this criterion always be fulfilled? No. We see a violation of it in the thermometer example above. An instance of the theory that fluids expand on heating, namely, that this mercury expands on heating, is entailed by the observation statement used above to test the theory, namely, that the temperature of the mercury thermometer is, say, 72° now, and was 62° five minutes ago (at times when its volume was greater and less, respectively).[30] This does not compromise the test. We have from the electrical resistance method independent evidence for the claim that this mercury expands on heating. It is sometimes possible to have such independent evidence despite relations of factual dependence between the things that a theory and observation statements used in testing that theory make claims about. We should reject this strong reading of the background independence criterion.

Consider now the epistemic (and intended) reading of Sober's criterion: it must not be the case that in order to know the observation statements used to test a theory, you have to believe that theory (or part of it). This is clearly an improvement over the previous, but it still need not always be fulfilled. What if we are once again testing the theory that fluids expand on heating, but have access to one of the fluids whose temperature we are measuring for only a brief time, during which the only instrument that interacts with it is the mercury thermometer? After losing access to the fluid and its condition forever, but retaining the thermometer and the reading we got from it, we realize that whether the reading of the thermometer reliably indicates temperature is not independent of whether the theory under test about the relation between temperature and volume is true. To believe that the thermometer reliably indicates temperature requires belief in part of the theory under test. Bad news, but we know what to do: find another method of measuring temperature that will assure us that the thermometer that had access to the fluid is a good measurer of temperature.

This is a situation where, among the beliefs that we have to have in order to know the temperature of the fluid, is a belief about the temperature and volume of the mercury thermometer. However, the belief about the mercury thermometer is not the basis for our belief about the temperature of the fluid. It is an epistemic conduit but not a provider of grounds for our belief about the fluid's temperature. And the other method of measuring temperature provides all the grounds we need. So this situation fails Sober's criterion but qualifies as a good test. Sober's criterion can be improved to withstand this example by taking 'x presupposes y' to mean that a belief in y is part of the basis for one's belief in x. In this form, I endorse the criterion in light of my discussion in §2 of why lack of candor is a problem, and my subsequent discussion of how the problem of a noncandid theory can be overcome, which uphold the idea that the theory's role as basis for beliefs about the reliability of outcomes must be eliminated by evidence that screens off the theory and is available independently of the theory. Thus I endorse the claim that we should disallow tests whose observational outcomes (including reliability) cannot be known without using beliefs about the theory under test as a *basis* for our beliefs about the outcomes and their significance. Furthermore, we have seen that probabilistic relevance keeps track of those basing relations remarkably faithfully.

Brown (1994) noted by an example that unfavorable outcomes are possible even when a theory has been assumed in the test of it as a basis for one's beliefs about the outcome of the test. However, Brown was mistaken in thinking that the issue about whether background dependence is acceptable is whether unfavorable outcomes are possible when the tested theory is assumed. Rather the issue is what any outcome of a test can tell us when the theory under test has had an influence on our construal of that outcome. The kind of example Brown and Hudson have in mind when they say that the possibility of unfavorable outcomes makes a circular test acceptable must be cases where the theory under test is either positively or negatively probabilistically relevant to the reliability of that unfavorable outcome. We saw above that neither kind of test can be revealing if there is no independent evidence for the reliability auxiliary.[31]

A possible counterexample to the stand I am taking against Brown and Hudson on this point may leap to mind. Did we not use Euclidean assumptions about the geometry of space to certify the working of the telescopes used to confirm Einstein's general theory of relativity—and thereby to refute the claim that physical space was Euclidean? Thus Euclidean geometry was a basis of our belief that the test indeed refuted the claim that geometry was Euclidean. All of this is (roughly) true, but this example does not favor Brown and Hudson. The reason we should believe that such a test was revealing is not that even when we gave the Euclidean theory an advantage by assuming it, that theory was refuted. Rather, the theory that needed to be assumed and the theory that was refuted were not the same. The theory that needed to be assumed was that Euclidean geometry is right at the scale of telescopes. The refuted theory is that Euclidean geometry correctly describes space at a much larger scale. Moreover, the theories of space in competition in these tests agreed that Euclidean geometry was correct at the scale of telescopes, and *that* is why using Euclidean geometry as a basis for our beliefs about the test is acceptable: we only needed to use the part on which the rivals agree.

The assumed theory, that Euclidean geometry is right at the local scale, was neither refuted nor in question in the test. Thus, although the theory under test, general relativity, makes the local Euclidean geometry used in the test probable, and so is relevant to it, nevertheless since the theory that says space is Euclidean at the large scale is relevant in the same direction to the same degree as general relativity, the assumption that Euclidean geometry is right enough on the local scale can safely be made in the test as if we had independent evidence for it. It is a legitimate question whether that assumption is true, but not a question that this test is supposed to decide. This illustrates that lack of candor of a given theory in a given test may fail to be a problem for that test in another way: the test may be able to distinguish rival theories if the two theories are relevant to the reliability of the outcome in the same way, to the same degree. A test may be able to distinguish the rivals if the matter about which the theories are not candid is not exactly the matter through which the test is designed to distinguish them.

The latest version of Sober's criterion, which I just endorsed, does not take a stand on what kind of relations of dependence among things in the world lead to situations where epistemic independence between observational test outcomes and the theory under test is lacking, but says that the latter is bad whatever has caused it. Consider next the requirement of background independence with a slightly different reading of the restriction against interdependence. This version of the requirement says that it should never be that the theory under test is probabilistically relevant to an auxiliary about the reliability of outcomes of the test, i.e., a theory must never fail to be candid for a test. We have seen that there are, in turn, two ways to read this requirement, one in which the probabilistic relevance of two statements is considered in isolation, and the other in which that relevance is judged taking all background knowledge into account.

On the first reading the requirement cannot be endorsed, contra, perhaps, Popper, since two statements may be probabilistically relevant when considered in isolation but not probabilistically relevant when independent evidence is taken into account. A theory may lack candor for a test in the sense that the theory is probabilistically relevant to the reliability of outcomes when the two are considered in isolation, but if independent evidence for the auxiliaries is acquired, that probabilistic relevance literally disappears. The other way of reading lack of candor, that it obtains only when the theory is probabilistically relevant to the reliability of outcomes *all things considered*, allows us to endorse a prohibition against noncandid theories, because a theory will fail to be candid in this sense only if there is no independent evidence for the auxiliaries. Both points cohere with the earlier argument that our default attitude toward theories lacking candor in a test should be suspicion, because if a theory lacks candor in either sense and there is no additional information, then there is no independent evidence, and the lack of candor has not been overcome. In this sense, Popper was right.

Advocates and critics of the requirement of background independence have both been both right and wrong. Critics have been wrong to conclude that "background independence is irrelevant to the prospect of gaining informative empirical data" (Hudson 1994, 605). As I have argued above, there is a serious epistemic problem if there is relevance between theory and (reliability) auxiliary and no independent evidence for the auxiliary. However, critics have been right to think that background dependence has been rejected too uniformly by methodologists, though the critics did not see the reason that they were right: background dependence does not make independent evidence for the auxiliary impossible, and it is that independent evidence that matters most. Advocates of background independence have been wrong to the extent that they issued a blanket rejection of probabilistic relevance between theory and test procedures, probably wrongly thinking that probabilistic relevance (in isolation) made independent evidence impossible. But they have had sound intuitions that their critics did not about the dangers inherent in circular testing, dangers that will be lurking whenever there is relevance between theory and reliability-auxiliaries and we do not know that we have independent evidence for the latter.

To relate these results to circularity more precisely, we need to distinguish two different kinds of circularity in an argument or test. We often say that an argument is circular when the arguer assumes what she is trying to prove. But this formulation is ambiguous as to whether the assumption is merely believed or is also playing the role of grounds for the conclusion in this argument. In the first case, which I will call *belief circularity*, one believes some part of the conclusion before executing the argument, but that belief does not play the role of grounds for the argument's conclusion. In the second, which I will call *justification circularity*, one not only believes some part of the conclusion before executing the argument, but that belief is being used in the argument as a basis for the conclusion. The second kind of circularity is obviously objectionable, whereas the first kind arguably is not. The notion of probabilistic relevance does very well in distinguishing the two, by means of the possibility of screening off relations, and because probabilistic relevance is not transitive.

When we speak of observations or outcomes of a test being laden with the theory under test, and we take this to be obviously objectionable, we are probably assuming that antecedent belief in the theory is producing justification circularity. Protesters would be right to respond that theory-ladenness might only cause belief circularity. But my argument above about what our default attitude should be toward a theory lacking candor in a test suggests that we should reply that the burden of proof is on the protester. It is not our burden to show that theory-ladenness is causing a problem, but his to show that it is not. He must show that the given test does not involve justification circularity, usually by showing that independent evidence for the auxiliaries has been acquired.

Franklin et al. (1989) challenged those who think that if observations are laden with the theory under test, that prevents a test of the theory. Could they come up with a workable example from actual science? Franklin et al. averred that they knew of none. They cited the thermometer case I have treated here as empirical evidence that theory-ladenness is not a problem, without explaining formally (or otherwise) why not. However, although they are right that this case does not present a problem, the thermometer case does not show that theory-ladenness with the theory under test is not a problem. Rather it shows that the problems introduced by theory-ladenness can in many cases be overcome by independent evidence.

More generally, the results here imply that if there are not examples in actual science of tests whose circularity thwarts them, this does not count as grounds for thinking that a situation where evidence is laden with the theory under test is unproblematic. The explanation of finding no examples in real science where circularity thwarts testing could instead be that the only cases of circular testing that survive to be recorded in accepted scientific venues are those in which investigators have found independent evidence to overcome the problem. The investigators would have thereby transformed a test from one that would have been justification circular to one that was merely belief circular. On this view the difference between some scientific theories and theories like psychoanalysis and Hegelianism would not be that one set of theories is honest and the other dishonest, but rather that though both sets possess a potential for "dishonestly" interfering with their own testing, for the first set we have found ways to overcome that problem and for the second less so.

To return to the analogy with candor in people, our situation with some theories in some tests is similar to the one in which we find ourselves when dealing with a person whom we suspect is not candid, and may even be lying. We have the option of pursuing evidence that does not depend on believing what the person says, and that is not restricted in content to the topics that the person has decided to discuss. If we are lucky, we find such evidence. Finding such evidence may also require ingenuity, but it will not come as news that testing scientific theories requires ingenuity.

ACKNOWLEDGEMENTS

I am grateful to Elliott Sober, Richard Grandy, Daniel Osherson, Hartry Field and John Roberts, and to Nicholas Asher, R. James Hankinson, Cory Juhl, Sahotra Sarkar, and other members of the Central Texas Philosophy of Science Consortium, for helpful discussions. I thank Louis Guenin for his patience with the editing of a long and cumbersome manuscript.

NOTES

[1] Note that there will be no problem if an empirical criterion can be given for distinguishing trustworthy from resisting denials. Psychoanalysis sympathizers tend to cite intensity of affect as a way of distinguishing a trustworthy denial from a resisting denial. (It is that Lady Hamlet protests *too much*, not that she protests at all, that gives her away.) On the other hand, could not intense affect in the denial be due to anger at the accusation (or felt incompetence) of the analyst, and so, be unrelated to repression?

[2] When I say that two statements are relevant intuitively I mean, for reasons that will become clear, to leave open whether they are probabilistically relevant in the technical sense.

[3] In making whether a theory lacks candor or not depend on its being probabilistically relevant to test procedures, I seem to be setting up a hair trigger for theories to be deemed 'bad.' Probabilistic relevance comes in degrees and may be very weak: should we count a case of weak relevance as lack of candor? It should be counted as lack of candor because that relevance, though weak, will have the structure I describe below. It is likely to be insignificant if the relevance is weak, and in any case I will be showing how lack of candor can be overcome by revealing tests, so that the property need not be bad all things considered.

[4] Of course, Hegel might not think that contradictions as described are necessary to criticism. (What about criticizing by offering an alternative explanation?) And it is not Hegel's or his theory's fault that Popper does think this. One may notice that Popper's argument implies that Hegel's theory protects from criticism not only itself but also every other theory that contains contradictions.

[5] Popper's charge just quoted is one of many examples where his accusations of reinforced dogmatism that seemed to indict a particular theory tended to become instead claims against the psychological dispositions of the theory's advocates, charges that could be appropriate against uncritical adherents of any theory. As I have said, such psychological or moral claims are not the interest of this paper.

[6] Sometimes propositions can be tested against their negations (e.g., 'Is Ms. Smith pregnant or not pregnant?'), but often they cannot, and I am assuming that in the latter cases which other theory a theory is being tested against is part of what forms the context defining the test to which the theory is being put. When I say that a theory is testable, then, I imply that there is a salient rival hypothesis that it can be tested against in some test.

[7] In this paper, to minimize complexity, I am going to be deliberately naïve about judgments of whether an outcome was (prima facie) favorable or not. This will focus attention on auxiliaries involved in judgments of reliability of test procedures. So I am assuming that from an outcome of an experiment we can know straightforwardly whether it was favorable or unfavorable to the theory. This is despite the fact that auxiliary assumptions are, of course, typically involved in those judgments too, because, as I have said, auxiliaries are involved in deriving a theory's predictions.

[8] The following nice example illustrating failure of transitivity of probabilistic relevance is due to Richard Grandy. Let A be 'lives in Houston,' B 'is on a Rice soccer team,' and C 'is female.' Assume that our population is people in the U.S., and that the male: female ratio in Houston is the same as it is in the entire U.S. Then A is (positively) relevant to B, B is (positively) relevant to C since, as it happens, $P(C/B)=1$, but A is not relevant (positively or negatively) to C, because someone's living in Houston does not change the probability of that person's being female from what it was by virtue of being a member of the U.S. population. The nontransitivity of probabilistic relevance can also be proven with a sandbox diagram. Draw a square to represent the universe of possibilities, and draw a horizontal line through the square to represent the separation of A's from not-A's. Draw a line through some point on the first line at a 45° angle (in either direction) from the first line, this to represent the border between the B's and the not-B's. Now draw a line through the intersection of the first two at a 45° angle from the second line in the same direction as the second line was measured in, this to represent the border between C's and not-C's. In this representation, two properties are probabilistically independent if and only if their border lines are orthogonal. Notice that though A is probabilistically relevant to B and B is probabilistically relevant to C, A is not probabilistically relevant to C. The border lines of A and C are orthogonal because $45^\circ + 45^\circ = 90^\circ$.

[9] The results below, and the responses we should have to them, would be similar if we took the auxiliary to be used to determine whether the result is relevant to testing the theory instead of whether the result is reliable. Some of my setup is similar to Hartmann 2001. (See also Bovens and Hartmann 2002.) I have generalized by adding a variable about the favorability of the outcome. Probabilistic relevance between theories and auxiliaries was also discussed in Roush 1999.

[10] Of course, H is relevant to H since everything is (maximally) relevant to itself, at least in the probabilistic sense of relevance, so there is a trivial sense in which H boosts its own probability. The question here is whether other relevance relations (via the auxiliary) are going to be such as to provide an additional route through which the prior probability of H will affect the posterior probability of H, so that the effect of the former on the latter will be illegitimate.

[11] I also assume that we do not know in the default situation whether the theory's rival in this test is probabilistically relevant to the auxiliaries in the same direction to the same degree as is the theory. If the rival is so relevant, then the test will not seem unfair from the point of view of either rival. Whether such a test seems revealing or not in an absolute sense will depend on the particular subject matter in question in the auxiliaries and whether the assumptions that the two theories make about it are plausible independently of them.

[12] At least, it follows on any view that takes positive relevance as necessary for evidence, since we are assuming that there is nothing relevant among our other beliefs.

[13] In what follows in this section I regard increase in probability automatically as confirmation and decrease in probability automatically as disconfirmation. In this respect the analysis depends on a notion of confirmation based on some relevance measure.

[14] My claims in these last few paragraphs about what would happen probabilistically are based on the obvious analogy of these circles to money pumps, which our beliefs will not be subject to if they conform to the probability calculus.

[15] Note that in developing these points I have not assumed that the theory contributes to any deviation from the following conditions: P(rel/fav. T) $\cong$ P(rel/T), and P(unrel/unfav T) $\cong$ P(unrel/T), where 'fav' means the outcome is favorable to the theory, and 'rel' means the outcome is reliable. That is, I have not assumed, for example, that it is more likely for the theory to count an outcome as reliable if it is favorable, or more likely to count an outcome as unreliable if it is unfavorable. This would be a particularly devious feature for a theory to have. It seems possible that psychoanalysis, under my cartoon description, has this feature, but according to the current result a theory need not have this feature in order to interfere with its own testability. The probabilistic relevance I cited was enough to derive the uninformativeness of tests.

[16] I am not the first to have noticed that probabilistic independence and independence of evidence are different notions, but there are several reasons why their difference deserves discussion here. One is that the distinction has often not been adhered to in discussion of realistic examples, as with the questions about testability that Popper raises or general matters of testability, in philosophy of science. Another is that it is only recently that attempts have been made to give formal accounts of the intuitive notion of independent evidence (see Sober 1989; Fitelson 2001). Finally, what has been done by way of a formal account until now has not treated the case where the hypothesis whose evidence is independent or not (from something) is an auxiliary, and where one of the things it should be independent from is the theory under test.

[17] (2a) should be read 'The probability of A given T and e is equal to the probability of A given e.' (2b) should be read 'The probability of e given A and T is equal to the probability of e given A.'

[18] Because if Z screens off Y from X then it follows that Z screens off X from Y, when Z screens off Y from X we can say 'Z screens off the relevance between Y and X.'

[19] 'Shielding' was Hans Reichenbach's original term for what we now call 'screening off.'

[20] Temperature can be inferred by measuring the electrical resistance of materials, or by measuring the intensity of the maximum wavelength of emission of radiation from a gas and knowing the identity of the gas, from which one can infer the temperature via a black-body curve. Of course, if these methods could be certified only by methods involving the increase of temperature with volume then a problem might remain. However, it seems that this is not so. (See what follows in the main text.) With these methods we may encounter the problem that we cannot measure length independently of temperature, and we must measure a length

in order to measure the temperature. However, this is a different problem, which seems to be solvable. See Gillies 1972, 18–20.

[21] It is plausible that this strongest possible form of relevance, where the theory *implies* the auxiliary, will be the least common kind of relevance in actual cases.

[22] In other cases e could turn out to be independent of T while being relevant to A, which is relevant to T, because relevance is not transitive.

[23] I am assuming that what is evidence for T will be evidence for A, which will not always be fulfilled when T is relevant to A, even if we assume a positive relevance notion of evidence, since relevance is not transitive. However, in cases where there is no evidence for T that is also evidence for A, the question we are now addressing will not arise, since any evidence for A will be distinct from the evidence we had for T.

[24] Note that in the special case where e' is T, which occurs because T is relevant to T, these conditions say that e is independent of T relative to A, which we also want to be true.

[25] Notice that I have not required that $\mathrm{P}(e/e') = \mathrm{P}(e)$, an absolute independence between the two pieces of evidence. In this I follow Sober's (1989) point that in many common conditions we should prefer conditional to unconditional independence between e and e' if we are looking to infer the state of a common cause of them. This builds on Reichenbach's earlier observation that joint effects of a common cause will not be unconditionally independent of each other (Reichenbach 1956, 159–160).

[26] Note that in case e is the only evidence we have—there is no e'—our intuition is that e provides the kind of evidence we need in order to trust our auxiliary in testing the theory. This is borne out in the second set of independence conditions (3) and (4) because if e' is erased from them, the e that I have described fulfills the conditions.

[27] Curiously, Hudson, who uses both of these ideas in defining background independence, accepts the requirement of background independence if it means an epistemic constraint (Hudson 2000, 6), but does not think that is what is meant by the phrases that he employs in the statement of the requirement of background independence, the requirement that he rejects. Since we are worried about epistemic grounds for drawing conclusions from tests, I do not see what else they should mean.

[28] This gloss on Sober's requirement comes from personal correspondence.

[29] Thus Peter Kosso is on the right track in attempting to define a notion of the independence of an account in his helpful (Kosso 1989). However see Roush 2004b for improvements on his notion of evidence that is independent of the theory under test.

[30] It may seem a gross abuse of language to call this an observation statement when as we have seen earlier our justification for believing it may depend on independent evidence about something as complicated as an electrical resistance measuring apparatus. However, a statement's being an 'observation statement' need not imply that it is known without recourse to other statements.

[31] For the same reason, a condition that Clark Glymour requires for evidence, namely, that "to test a hypothesis we must do something that could result in presumptive evidence against the hypothesis," is necessary but not sufficient for evidence that is independent of the theory under test (Glymour 1980, 115).

REFERENCES

Achinstein, P.: 2001, *The Book of Evidence*, Oxford University Press, Oxford.

Bovens, L. and Hartmann, S.: 2002, 'Bayesian Networks and the Problem of Unreliable Instruments', *Philosophy of Science* **69**, 29–72.

Brown, H.: 1994, 'Circular Justifications', *Proceedings of the Biennial Meeting of the Philosophy of Science Association 1994*, vol. 1, Contributed Papers, pp. 406–414.

Carnap, R.: 1950, *Logical Foundations of Probability*, University of Chicago Press, Chicago.

Copernicus, N.: 1952, *On the Revolutions of the Heavenly Spheres*, Encyclopedia Brittanica, Inc., Chicago.

Feyerabend, P. K.: 1993, *Against Method*, Verso, New York.

Fitelson, B.: 2001, 'A Bayesian Account of Independent Evidence with Applications', *Philosophy of Science* **68**, S123–S140.

Franklin, A. et al.: 1989, 'Can a Theory-Laden Observation Test the Theory?', *British Journal for the Philosophy of Science* **40**, 229–231.

Gillies, D.: 1972, 'Operationalism', *Synthese* **25**, 1–24.

Glymour, C.: 1980, *Theory and Evidence*, Princeton University Press, Princeton.

Grünbaum, A.: 1984, *The Foundations of Psychoanalysis: A Philosophical Critique*, University of California Press, Berkeley.

Hartmann, S.: 2001, 'The Import of Auxiliary Theories of the Instrument: A Bayesian Approach', *APA Pacific Division Meeting*, San Francisco, CA.

Hudson, R.: 1994, 'Background Independence and the Causation of Observations', *Studies in History and Philosophy of Science* **25**, 595–612.

Hudson, R.: 2000, 'Evaluating Background Independence', Philosophy of Science Association Biennial Meeting, Vancouver, B. C.

Kosso, P.: 1989, 'Science and Objectivity', *Journal of Philosophy* **86**, 245–257.

Kuhn, T. S.: 1996, *The Structure of Scientific Revolutions*, 3rd. ed., University of Chicago Press, Chicago.

Lakatos, I.: 1978, *The Methodology of Scientific Research Programmes*, ed. Worrall, J. and Currie, G., Cambridge University Press, Cambridge.

Popper, K.: 1966, *The Open Society and Its Enemies*, vol. 2, *The High Tide of Prophecy: Hegel, Marx, and the Aftermath*, Princeton University Press, Princeton.

Popper, K.: 1989, *Conjectures and Refutations: The Growth of Scientific Knowledge*, Routledge & Kegan Paul, London.

Reichenbach, H.: 1956, *The Direction of Time*, University of California Press, Berkeley.

Roush, S.: 1999, *Conditions of Knowledge*, Ph. D. dissertation, Harvard University.

Roush, S.: 2004a, 'Discussion: Positive Relevance Defended', *Philosophy of Science* **71**, 110–116.

Roush, S.: 2004b, 'Testability and the Unity of Science', *Journal of Philosophy* **101**, 555–573.

Salmon, W.: 1975, 'Confirmation and Relevance', in Maxwell, G. and Anderson, R., eds., *Minnesota Studies in the Philosophy of Science*, vol. 6: *Induction, Probability, and Confirmation*, University of Minnesota Press, Minneapolis, pp. 3–36.

Sober, E.: 1989, 'Independent Evidence About a Common Cause', *Philosophy of Science* **56**, 275–287.

Sober, E.: 1999, 'Testability', *Proceedings and Addresses of the American Philosophical Association* **73**(2), 47–76.

Department of Philosophy
Rice University
Houston, TX 77005
U.S.A.
E-mail: roush@rice.edu