Episteme (2017) page 1 of 19 © Cambridge University Press
doi:10.1017/epi.2017.3

WHY NOT ALL EVIDENCE IS SCIENTIFIC EVIDENCE

CARLOS SANTANA
c.santana@utah.edu

ABSTRACT

Data which constitute satisfactory evidence in other contexts are sometimes not treated as valid evidence in the context of scientific confirmation. I give a justificatory explanation of this fact, appealing to the incentives, biases, and social situatedness of scientists.

1. A PUZZLE ABOUT SCIENTIFIC EPISTEMOLOGY

Norms of modern scientific practice include standards of evidence much stricter than the standards of evidence for everyday inference. My aim is to justify these stricter norms, even though they exclude some otherwise legitimate evidence from scientific consideration. Before getting into *why* not all evidence is scientific evidence, let's consider a few vignettes to convince ourselves that not all evidence is scientific evidence.

Vignette 1a

Quinn gets a nasty case of food poisoning. She calls her sister, who recommends boiled cloves. 'I was skeptical when I first heard about it,' her sister says, 'but I tried it and it did wonders for nausea.' Even though she's not fully convinced it will be effective, on the basis of her sister's experience Quinn gives the folk remedy a try.

Vignette 1b

Quinn is a pharmacologist trying to discover an agent that will reduce symptoms of nausea. She calls her sister, who recommends boiled cloves. Quinn writes up a proposal for a NIH grant, noting that "We have some preliminary evidence that cloves reduce nausea, in the form of my sister's experience." After getting funding, and performing randomized controlled trials, Quinn applies for FDA approval to market cloves as an anti-nausea agent. On the application she cites all her evidence, including both the outcome of the RCTs, and her sister's anecdote.

It would be surprising to see an appeal to one individual's experience on a biomedical document. Nevertheless, it seems perfectly reasonable for Quinn to allow her sister's testimony to increase her confidence that cloves are an efficacious remedy. The example works as well with conventional wisdom substituted for anecdote: Quinn knows that it is conventional wisdom that cloves fight nausea, etc. In individual decision making, we

rely on common sense and conventional wisdom all the time, and this seems appropriate. Even though conventional wisdom is often wrong, following its dictates is probably more reliable than choosing randomly. Nevertheless, we would be surprised to see explicit appeals to conventional wisdom in most scientific writing.

Vignette 2a

Tevita is planning a vacation to the Amazonian Andes. He wants to get along with the locals, so he chats with a friend who had visited Peru the year before, and asks questions about the Quechua people. On the basis of what his friend says, Tevita forms several beliefs about Quechua customs.

Vignette 2b

Tevita is an anthropologist, writing an ethnography of the Quechua people. Money's tight, so rather than doing field research himself, Tevita interviews American tourists returning from Peru about Quechua customs. He uses these interviews as the primary data for his ethnography.

Tevita the scientist's actions seem wrong, while Tevita the tourist seems justified. Testimony is an ineliminable source of evidence in everyday life, but it seems that scientific norms exclude the testimony of non-experts in certain circumstances.[1]

Vignette 3a

Mariana is working on her calculus homework. She has little tolerance for working through problem sets but she's mathematically gifted. She often looks at a problem and just has a hunch about the answer, and these hunches are almost always right. She uses these hunches to fill out her homework instead of working through the problems.

Vignette 3b

Mariana is a gifted mathematician. She often has hunches of the form "p is a theorem of formal system q." It's well known that her hunches are generally accurate. Since this is the case, and since she hates coming up with formal proofs, her published papers are all of the form "I have a hunch that p, therefore p."

In personal decision making it makes a lot of sense to rely on hunches, intuitions, etc. These data are rarely acceptable as evidence in scientific discourse, however. They may play other roles, such as helping the scientist generate hypotheses, but do not generally count as evidence.

We could adduce examples from nearly every domain of science. Unlike the lay public, linguists don't consult high school English textbooks to determine grammaticality,

1 Note that who counts as an expert is determined by the question we're asking, not merely by possession of some credential. The testimony of American tourists is not scientific evidence for an ethnography of the Quechua, but it is evidence for an ethnography of American tourism.

geographers don't use automobile odometers to determine distance, and meteorologists don't rely on old sailor's rhymes to predict tomorrow's weather. Nevertheless, each of these sources of evidence seems safe for personal use.

These vignettes give us a feel for the puzzle we're going to tackle. In each of these cases, a kind of data which constitutes satisfactory, though perhaps weak, evidence for an individual epistemic agent does not constitute satisfactory evidence for the joint epistemic project of science.[2] So, in asking "why isn't all evidence scientific evidence?" we're looking for the features of scientific epistemology which make it relevantly distinct from individual epistemology.

Since I'm claiming that not all evidence in the strict sense counts as evidence in science, we need a definition of scientific evidence.

SCIENTIFIC EVIDENCE: *E* is scientific evidence for a hypothesis *H* iff *E* is evidence[3] and *E* is acceptable as confirmatory in scientific practice.

I intend 'confirmatory' to mean incremental, not just absolute confirmation (cf. Carnap 1962), and it includes evidence which disconfirms as well as that which confirms *H*. The word 'acceptable' in this definition should be taken as normative, meaning that *E* is scientific evidence only if it's the kind of thing scientists should accept as evidence. With this definition in hand, we can see that evidence in the strict sense and scientific evidence are conceptually distinct, since a datum can be technically confirmatory without being acceptable as confirmatory in scientific practice. The vignettes illustrate how they are frequently actually distinct. Contemporary scientific practice excludes some data which is evidence in the strict sense. Anecdata, common sense, hunches, non-expert testimony, outdated measurement techniques, high school textbooks and so on are often evidence in the strict sense, but not scientific evidence.

One obvious reason why not all evidence is scientific evidence is that scientists have limited resources. If we only have the time or funding to pursue a limited number of observations, it makes sense to pursue the most epistemically valuable data. Weaker evidence might end up not being scientific evidence because we don't have the resources to gather it under scientific conditions. Most exclusions of evidence in the strict sense from scientific argumentation, however, can't be chalked up to limited resources, because in most cases the weaker types of evidence are already available or obtainable at very low cost. The scientist often already knows the conventional wisdom or anecdata that apply to the question at hand, and intuition-pumping, consulting non-expert opinion, and so on are all nearly costless means of data gathering. If anything, the effect of limited resources would be to favor weak evidence, so we can't explain why not all evidence is scientific evidence by a mere appeal to limited resources.

To make the puzzle even more precise, here's one more definition:

2 Note that this issue here is different from the one raised by Achinstein (1995, 2001), that not everything which raises the probability of a hypothesis intuitively counts as evidence. In the cases I'm dealing with, it does make sense to call the data evidence, just not in scientific contexts.

3 I'm deliberately non-committal on what it is to be evidence in the strict sense. I assume that any reasonable account of evidence (that evidence is the thing which justifies belief, that evidence = knowledge, that evidence is that which should alter credences, etc.) will suffice to generate the puzzle, because whichever of these definitions we accept, some evidence in the strict sense will be excluded by contemporary scientific practice.

PRINCIPLE OF TOTAL EVIDENCE (PTE): In decision making and hypothesis confirmation take into account all available relevant evidence.

The name comes from Carnap (1947), but the idea is an old one. Just to be clear, PTE doesn't say that before you make a judgment you must go out into the world and track down every piece of relevant data first. It merely recommends using all the data you have available at the time of judgment, as well as any that can be obtained costlessly. Carnap justifies PTE in part by noting that it is "generally recognized" and that it would be "obviously wrong" to violate it (1947: 138–9), and certainly PTE is quite intuitive. I.J. Good, however, proposes that PTE is not supported by intuition alone, arguing that PTE follows from "the principle of rationality – the recommendation to maximize expected utility" (1967: 319). Good makes his case using a mathematical proof, but the basic idea is understandable qualitatively. When making a decision, taking into account an additional piece of evidence can only increase the objective likelihood of achieving the best available outcome. So as long as using or obtaining an additional piece of evidence is costless, it is always beneficial from the standpoint of utility to do so. Since an individual's total evidence includes only that which they already possess, and evidence already-in-hand is costless, to maximize expected utility an individual must follow PTE.

If you don't think that questions of epistemic rationality can be treated by appeal to practical rationality and expected utility, you'll need grounds other than Good's to accept the principle of total evidence. Perhaps Carnap's appeal to its intuitiveness works for you; perhaps you have your own reasons. Whatever your reasons, if you accept PTE,[4] we've come to the puzzle. PTE states that epistemic rationality demands that we take into account all our evidence. But in scientific practice we don't do so (or at least pretend not to), and in fact we seem to think it is epistemically wrong to do so. So either scientific practice egregiously violates a straightforward epistemic principle, or PTE doesn't apply in scientific contexts. I favor the latter for two reasons. First, PTE doesn't apply to scientists because it is a principle for ideally epistemically rational agents, and scientists are not much more ideally rational than the rest of us. Second, PTE is meant to apply to individual agents, but scientific inquiry is inherently social.

These two facts mean that the epistemic ends of science are better served by a different principle of evidence, which we'll call the scientific standard of evidence. Different kinds of evidence vary in what I'll term *reliability*, where a token piece of evidence from a more reliable type of evidence is less likely to be misleading than a token piece from a less reliable type. The scientific standard of evidence states that, roughly, only the most reliable sources of evidence are acceptable as scientific evidence. We can make this more explicit.

SCIENTIFIC STANDARD OF EVIDENCE (SSE): A type of evidence is acceptable scientific evidence if either

(a) It is highly reliable, or
(b) It is among the most reliable types of evidence available.

Let's call condition (a) the absolute criterion and (b) the relative criterion.

4 If you reject PTE, perhaps you accept a near neighbor which yields the same puzzle. If you reject anything in the ballpark, then there is no puzzle, and you can treat the puzzle-solving I'm about to engage in as further arguments for a claim you already accept – that PTE is false.

The idea behind the absolute criterion is that we don't always want to exclude very good evidence just because nearly impeccable evidence is available. What counts as highly reliable probably varies a bit from discipline to discipline, but all that matters is that a vague threshold for absolute reliability exists. The relative criterion exists because for some scientific questions good evidence is hard to come by. This doesn't mean we give up on asking those questions, it just means that we have to make do with the best evidence available. Having a relative criterion as a disjunct to the absolute criterion ensures that we can do science even when the only evidence available is poor.

To apply the definition of SSE, we also need a means of individuating types of evidence. My suggestion is that we take two different pieces of evidence to belong to the same type when they are gathered, measured, and assessed using the same method. A fingerprint analysis and a DNA analysis, for instance, are evidence for the same thing – the identity of the suspect – but are gathered and processed using different methods, so they belong to separate evidence types. DNA analysis of a skin sample, however, might produce evidence of the same type as DNA analysis of a blood sample, unless they are gathered or assessed using a different method. I suggest this means of individuating types of evidence because it carves the world of evidence at the right joints: between sets of evidence that are roughly similar in terms of reliability. It may be that other methods of individuating evidence types accomplish the same thing, and if so, they will be just as compatible with my arguments in favor of SSE.

A few further comments on SSE. First, SSE is an aspirational norm and mechanisms for enforcing it are often informal, imprecise, and indirect, so we should expect to see imperfect adherence to it. Occasional violations of SSE don't demonstrate its inexistence. It does seem, however, that SSE accurately describes the standard of evidence across a wide array of sciences – consider the variety among the vignettes with which we began. Second, successful application of SSE presupposes that we have at least a rough idea of how reliable a type of evidence is.[5] So where the level of reliability is in question, SSE predicts that we should expect scientists to dispute whether a certain type of evidence is acceptable. Likewise, SSE explains in part why reproducibility is so important – it helps establish the level of reliability of a type of evidence.

Having specified SSE, we can restate the puzzle as "Why does science use SSE rather than PTE?" The answer, as I've suggested, is that science is a collective endeavor of non-ideal agents. To demonstrate, I'll compare what happens if science follows SSE to what happens if it follows PTE, showing that in many common situations the epistemically superior outcome comes from following the stricter criteria in SSE. I'll use case studies and simulations to show that these counter-intuitive results fall out of scientists' bounded rationality and the social nature of science. Showing that SSE leads to superior epistemic outcomes than PTE is sufficient to justify the scientific departures from PTE.[6] It shows, in short, that treating relatively weak evidence as counting as evidence in science will

5 Several commentators have suggested to me an even stronger prerequisite – that we know *why* the type of evidence in question is reliable. I think this is descriptively inaccurate, but won't argue the point here, since nothing in this paper hangs on whether or not we accept this stronger requirement.

6 Less directly, it may also help *explain* the sociological fact that PTE is generally not the standard of evidence in science. Here's one speculative possibility: research programs that adhere to SSE are more successful than programs that adhere to PTE, for all the reasons presented in this paper. Researchers in those programs thus attract more students and funding, and other research programs imitate their norms, both of which lead to the propagation of the SSE norm.

frustrate the goals of science. We thus have pragmatic reason to think of evidence in a different way in the scientific domain than in other domains.

2. EVIDENCE AND INCENTIVE

Let's idealize for the moment and stipulate that the goal of science is to come up with true theories. The first reason to prefer SSE to PTE is that the goal of the individual scientist is different. Although most scientists aim to produce true theories, they also have other ends, including receiving credit for their work and using their time efficiently.

The chief method by which a scientist gets credit is by publishing research. Generally, these publications take the form of a set of claims and the evidential support for this set of claims. Consider a scientist aiming to publish a paper she has written arguing for hypothesis *H*. Imagine that the norms of her discipline accept two sorts of evidence: *e*, which is weak but easy to obtain, and *E* which is strong but costly to obtain. Knowing that her paper will be accepted even if containing only evidence of type *e*, she will likely only try to obtain evidence of type *e*. Her peers will generally do the same, because it's the most efficient means to gain credit. In the aggregate, this leads to suboptimal epistemic outcomes, since the better sort of evidence *E* will be underutilized in the discipline.

Note that the problem is less severe in the case of the individual agent. An individual's decision about whether or not to pursue *e* or *E* will be determined by weighing the cost of obtaining *E* against the benefit it provides for their expected utility. If obtaining *E* is worthwhile, the rational agent will do so. In the case of a group of scientists, however, even if all the scientists are ideally rational they may all individually choose *e* because it dominates choosing *E* from the individual perspective. This of course undermines the goal of their scientific discipline. Depending on how much the scientists themselves value obtaining true theories, it may also be sub-optimal for their own utility, in a sort of social dilemma.

Let me give two examples. I'm hesitant to make critical claims about a specific scientific discipline without detailed, careful argumentation, so both examples will be of scientists criticizing their own discipline on the grounds that their colleagues have flocked to weak evidence *e* at the cost of strong evidence *E*.

Our first example comes from generative syntax. Wasow, a syntactician, and Arnold, a psycholinguist, make the following claims about evidence in generative syntax: "standards of data collection and analysis that are taken for granted in neighboring fields are widely ignored by many linguists. In particular, intuitions have been tacitly granted a privileged position in generative grammar. The result has been the construction of elaborate theoretical edifices supported by disturbingly shaky empirical evidence" (Wasow and Arnold 2005: 1481–2). This is a polemical claim, of course, so Wasow and Arnold back it up with a detailed argument. Their argument, in fact, proceeds just as we would expect it to if SSE were taken as normative in linguistic research. First they outline the varieties of evidence available to the syntactician, including intuitions, corpus data, and psychological experiments. They then give evidence that intuitions are markedly less reliable as evidence than the other two sources. Intuitions, to use our present terminology, are *e*, while experimental and corpus data are *E*. Among the evidence they cite for this claim are empirical results showing that intuition data exhibits much more variability than usage data. Moreover, they illustrate with a number of examples that when a linguist's intuition runs counter to usage data, we reject the intuition in favor of the usage data, which shows that usage data is accepted as the more reliable sort of evidence. Their

examples are also meant to show that usage data contradict intuitions fairly frequently, so intuitions must not be highly reliable in an absolute sense. Taken together, these observations show that intuitions are (a) not highly reliable, and (b) significantly less reliable than feasible alternative sources of evidence. Wasow and Arnold note that despite this fact, most syntax papers appeal largely or only to intuitions as evidence, and they argue that this must be harmful to the discipline as a whole.

Let's assume that Wasow and Arnold are correct in their assessment of the uses of evidence in generative syntax. The problem seems to be that intuitions are both easy to obtain and considered satisfactory evidence by the field, so linguists are disincentivized from pursuing stronger evidence. This, unfortunately, diminishes the epistemic quality of the output of the field. Were the field to hew more closely to SSE, however, intuition data would no longer be sufficient to support a theoretical claim, so linguists would have to rely on stronger forms of evidence. What I'm suggesting, then, is that if Wasow and Arnold are right, generative syntax could improve the epistemic quality of its output by adhering to SSE. This seems to be what Wasow and Arnold think as well: they conclude by arguing that "linguistic inquiry should be subject to the methodological constraints typical of all scientific work" (2005: 1495). SSE, I'm arguing, is a reasonable account of those pan-scientific methodological constraints. If syntacticians accepted SSE, however, it would require rejecting PTE, since intuition-data is virtually costless.[7] In short, the case of generative syntax shows that the fact that scientists seek easy credit requires us to reject PTE in favor of SSE.

A similar example in a different discipline comes from the provocatively titled "Psychology as the science of self-reports and finger movements: Whatever happened to actual behavior?" (Baumeister *et al.* 2007). Consider two types of evidence available to social psychologists, direct observation of the behavior in question and self-reports. Call the first behavioral methods, and the second survey methods. Behavioral methods give us significantly more reliable evidence, so behavioral methods give us evidence E, while survey methods give us evidence e. This is not to say that surveys and self-reports are no better than uninformed guessing. They often are, and so constitute evidence in the strict sense, but they are much less reliable than behavioral methods. Unfortunately, they are also much cheaper. Surveys take less time, both to design and deploy, and are often easier to analyze as well. So if my point about incentive for easy credit leading to poor epistemic outcomes is true, we would expect a disproportionate number of social psychologists to be publishing survey evidence instead of behavioral evidence. According to Baumeister *et al.* we do indeed find this to be the case. They recount that

> personality psychology has long relied heavily on questionnaires in lieu of behavioral observation, a state of affairs that has begun to change only recently and ever so slowly, at that. Even worse, social psychology has actually moved in the opposite direction. At one time focused on direct observations of behaviors that were both fascinating and important – a focus that attracted many researchers to the field in the first place – social psychology has turned in recent years to the study of reaction times and questionnaire responses. These techniques, which promised to

7 Note that if intuition data were not virtually costless, then PTE would be compatible with Wasow and Arnold's conclusion, because PTE does not tell against prioritizing which methods we use to gather evidence. Even, however, in cases where weaker evidence is not virtually costless, a flat-out prohibition on weak evidence (SSE) will lead to the epistemically preferable outcome because it precludes the possibility of rationalizing the pursuit of weak evidence at the expense of strong evidence.

> help to explain behavior, appear instead to have largely supplanted it. The result is that current research in social and personality psychology pays remarkably little attention to the important things that people do. (Baumeister *et al.* 2007: 396)

Of course, the authors give evidence for this claim. They examined the two most recent issues of the reputable *Journal of Personality and Social Psychology*. In the most recent issue, they identified 38 published studies. Of the 38, only one involved direct observation of behavior, and that one they classified as borderline. The previous issue, however, was 100% better. Out of 38 studies in that issue, two involved direct behavioral evidence.

Note that the problem isn't that questionnaires aren't evidence. The authors of this complaint acknowledge that "Self-reports are often illuminating" especially because in some cases, such as studies of emotional experience, they "are all that is possible" (Baumeister *et al.* 2007: 399). In other words, when it comes to questions where the relative criterion obtains – where self-reports are the best option despite being unreliable in an absolute sense – psychologists should use self-reports. This stance coheres with SSE. No, the problem is that on all the questions where the relative criterion doesn't obtain the best available evidence is being neglected because lesser evidence is good enough to get credit. The solution, again, is to enforce something like SSE. In an ideal world where scientists were incentivized merely to provide the best epistemic contribution possible, they would find the best balance of gathering both behavioral and survey data. But scientists are not so incentivized. If we were to forbid appeal to self-reports when better data is available, it would lead to the use of better evidence and thus truer theories in social psychology. Self-reports would join anecdata and appeals to conventional wisdom in the bin of weak evidence not worth our time. By adopting SSE, we would improve the epistemic quality of the science.

I don't mean to beat up on social psychology or generative syntax. Both fields produce interesting results, and many researchers do avail themselves of the best techniques. The point is merely that leading researchers in both fields have identified a problem caused by credit-seeking and claim it could be resolved by adherence to something like SSE. I don't mean to beat up on credit-seeking either. It plays an ineliminable role in science, since we want motivated scientists. SSE, unlike PTE, allows us to both encourage credit-seeking and mitigate the potential it has to epistemically undermine scientific practice. So one reason for favoring SSE, attested in at least two contemporary disciplines, is that it helps ensure that the best sources of evidence are not neglected.

3. COGNITIVE BIAS AND EVIDENCE-GATHERING

A second set of reasons why scientists should avoid PTE comes from the fact that scientists are as subject to cognitive biases as the rest of us. PTE assumes ideal epistemic agents, but common cognitive biases mean that none of us is ideal in the appropriate sense.[8] Consequently, following PTE can actually lead to worse epistemic outcomes in a number of circumstances. In this section, I'll review some well-demonstrated traits of human

8 Ideal, that is, in the sense of being the sort of agent in idealized decision-theoretic models. It has been argued on empirical grounds (Gigerenzer and Brighton 2009) that given our actual environment, such agents are not actually optimal reasoners. If this is the case, then the 'biases' discussed in this section are in fact good reasoning strategies for individual learners and the problems I attribute to them will emerge particularly in the context of joint scientific research.

reasoning, and show how given these traits, following SSE would lead to epistemically superior outcomes.

The first problem caused by cognitive biases has to do with the timeframe of evidence acquisition. Remember that at issue is whether scientists should ignore evidence that they get for free – in other words, the type of evidence at issue is the kind the scientist will already have been exposed to at the beginning of the inquiry. If they had not already been exposed to the evidence, they would have to seek it out, so it wouldn't be free, and PTE wouldn't apply. Many of the examples of questionable scientific evidence are of this costless, already-in-hand sort. Anecdotal evidence, intuitions, conventional wisdom, and so on all fall into this category. Furthermore, the types of data that don't are usually easier to gather, so if a scientist is going to gather multiple kinds of data, they usually start with the weaker but easier to obtain evidence, such as survey data.

The problem with this is that a slew of well-demonstrated cognitive biases show that there is a first-mover advantage in evidence assessment. One set of these biases has to do with overrating the first bit of evidence you receive. The bias that Tversky and Kahneman (1974) identify as the anchoring effect, for instance, is our tendency to do precisely that. They give as a simple example of anchoring the following experiment: two groups of high school students were given five seconds to estimate an arithmetic expression. One group was given $8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1$, and the other given $1 \times 2 \times 3 \times 4 \times 5 \times 6 \times 7 \times 8$. The two expressions both multiply to 40,320, but the median estimate for the first group was 2250, contrasting with a median of 512 for the students given the ascending sequence. Which evidence the students processed first, that is, had a disproportionate impact on their assessment.

The anchoring effect is similar to another cognitive bias called the focusing illusion. Focusing occurs when any piece of information salient in the immediate context is taken to be especially relevant to a judgment, whether it actually is or not. Strack *et al.* (1988) provide a clear experimental example of focusing. They asked subjects two questions:

(1) "How happy are you with life in general?" and
(2) "How many dates did you have last month?"

Asked in that order, answers to the two questions showed no significant correlation. When (2) preceded (1), however, correlation rose to 0.66, a very significant correlation for psychology. In other words, merely making a certain set of otherwise irrelevant facts salient made those facts dominate subjects' judgments. Both anchoring and focusing have been demonstrated repeatedly and shown to apply in a number of different situations (Kahneman *et al.* 2006). Furthermore, subjects do not appear to be able to avoid the biases even if they are made aware of them (Wilson *et al.* 1996) and have difficulty resisting the pull of anchors even when given monetary incentive to avoid them (Simmons *et al.* 2010).

So between anchoring and focusing we have two well-demonstrated and recalcitrant cognitive biases which make it so that we will tend to be disproportionately influenced by the first piece of evidence we attend to. Why is this a problem? For the following reason. These biases mean that we're going to overrate evidence no matter what. The question is how to mitigate the effect of this overrating. If we're going to take a piece of evidence to be more decisive than it actually is, better to go with a highly reliable

datum than a fairly unreliable one. Therefore, overrating weak evidence is more harmful than overrating strong evidence.

Suppose we use PTE in science. Then the first-mover advantage goes to the weak evidence. Suppose we really internalize SSE, however. The scientist will dismiss the weak evidence, and the first salient evidence will be the strong evidence. So with PTE we overrate and focus on weak evidence, but with SSE we overrate and focus on strong evidence. This gives us one reason to prefer SSE to PTE.

The first-mover problem caused by cognitive bias occurs in a different way as well. Suppose for the sake of argument that we have a scientist who successfully suppresses anchoring and related biases, and thus accurately updates her belief when presented with weak evidence. Will taking weak evidence into account still impede her scientific research? Yes, in many cases it will, because of motivated reasoning in the form of confirmation bias and research bias.

Confirmation bias, perhaps the most discussed, experimented-upon, and well-demonstrated human cognitive bias, is "unwitting selectivity in the acquisition and use of evidence" to support a favored hypothesis (Nickerson 1998: 175). This occurs through favoring confirmatory evidence, ignoring alternative explanations, looking only for positive cases, and over-weighting confirmatory evidence (Nickerson 1998). Related is research bias, also called experimenter bias, which is a catch-all term referring to the subtle and unintentional ways in which researchers' prior beliefs lead them to manipulate the outcome of experiments. Philosophers are familiar with research bias in the famous example of Clever Hans, a horse supposedly able to perform arithmetic, but who was actually responding to unconscious bodily cues from its trainer (Pfungst 1911). Clever Hans appeared to be able to do arithmetic only because its trainer believed that it could and thus unwittingly manipulated a demonstration of that fact. A more apposite and contemporary example is the frequency of p-hacking in psychological research. P-hacking occurs when psychologists take advantage of "researcher degrees of freedom" to make what should be a null result come out as a positive one (Simmons *et al.* 2011). For example, a researcher might add subjects until they find a positive result, then end the experiment. Similarly, they might (unconsciously) choose which statistical methods to use, variables to analyze, or comparisons to make in order to increase their chances of confirming their hypothesis. Research bias in the guise of p-hacking appears to be ubiquitous in published psychological research (Simonsohn *et al.* 2014), and since there's nothing special about the brains of psychologists, we should expect research bias to be common in other disciplines as well.

My intent in bringing up confirmation bias and research bias is not to cast doubt on scientific results, but to highlight a pitfall in research – a pitfall that can be exacerbated by adherence to PTE. Scientists following PTE will begin by consulting weak evidence, and update credence in their hypotheses on the basis of this weak evidence. Since the weak evidence likely agrees with their prior hypotheses, their credence in those hypotheses will go up. Likewise, if a scientist is open-minded, consulting weak evidence will shift him from his neutral stance. In either case, the consequence of consulting weak evidence is that the scientist is more likely than before to engage in confirmation-bias-induced shenanigans, tainting the epistemic outcome of their research. A scientist who adheres to SSE, on the other hand, will not have this increased chance of unwittingly manipulating results.

SSE then, can lead to better outcomes than PTE because it mitigates some of the problems caused by confirmation and research bias. We should note, however, that this is not

the case if the epistemic disvalue of the bias is outweighed by the epistemic contribution of the weak evidence. But this is generally going to be consistent with SSE. Recall that SSE doesn't proscribe comparatively weaker evidence if that evidence is fairly strong in its own right. Cases where the epistemic value of the weaker evidence trumps the negative effects of bias will probably be cases where the comparatively weaker evidence is strong enough to pass SSE. In other words, in either possible case, SSE performs as well as or better than PTE.

One response to this argument would be to argue that cognitive biases in science are too trivial to worry about. Because the errors introduced through biases such as confirmation bias are generally small and subtle, we might think that they are generally drowned out by good scientific methodology. If this were the case, we could safely ignore the cognitive biases of scientists. Unfortunately, this is not the case. Mathematical modeling shows that even small amounts of bias have a major deleterious effect on the percentage of scientific claims which are true (Ioannidis 2005). Additionally, most professional incentives in science probably increase confirmation bias (Nosek *et al.* 2012), meaning that the bias is unlikely to be small in the first place. Strict standards of evidence are thus necessary to mitigate some of this bias.

To recap, scientists are human, and they must contend with the effects of cognitive bias on their research. PTE does nothing to correct for these biases, but the hard line drawn by SSE does diminish their effect in common cases. This gives us additional reason to favor SSE, and perhaps an additional explanation for why scientists treat weak evidence as if it were not evidence at all.

4. EVIDENCE AND THE SOCIAL STRUCTURE OF SCIENCE

A third reason for favoring SSE over PTE comes from the collective nature of the scientific enterprise. In brief, the problem is that in situations where an individual determines their beliefs in part by taking the beliefs of their peers as evidence, weak sources of evidence can lead to a harmful false consensus. This is because if a number of researchers respond to weak evidence, it might create a quick consensus in the discipline, and this consensus can be taken to be strong evidence that a hypothesis is true. For example, suppose astronomers attend to the fact that folk wisdom suggests that the moon is made of cheese. Further suppose that each individual astronomer recognizes that folk wisdom is only weak evidence, and thus comes to believe the hypothesis that the moon is made of cheese only tentatively. So far so good; everything each astronomer has done so far is reasonable. At this point, however, each will notice that the hypothesis is nearly universally held to be true among their peers, but it will not always be plain that a peer's belief is weak or that it is based on weak evidence. In virtue of this, there will appear to be a consensus in the discipline that the moon actually is made of cheese, and this will solidify each astronomer's belief in the hypothesis. But if the folk wisdom serving as evidence has the potential to be systematically misleading, then this solidified consensus is unwarranted.

This is an instance of what economists call an information cascade (Banerjee 1992). In situations where you have reason to believe that others may have information you lack, it can be rational to follow the crowd, even if your private information contradicts the crowd's behavior. Behaviors can therefore spread through a population principally on the basis of their popularity, as has been confirmed in a number of empirical domains.

Voters, for instance are known to be influenced by opinion polls, and marketers know that they can create real demand for a product by buying up product to make it appear that such demand already exists (Bikhchandani *et al.* 1998). Although using popular behavior as a source of information in this way is often a sensible move from an individual perspective, it can easily lead to suboptimal group outcomes (Banerjee 1992), as in the moon-cheese case.

The reason the possibility of information cascades favors SSE over PTE is that information cascades are only likely to lead to false consensus if the evidence underlying individuals' original behavior is relatively weak. False consensuses are a particularly bad outcome for a scientific discipline. Not only can a false consensus directly obscure the truth, but future work premised on a false theory will also be misleading to the community. Additionally, false consensuses in science often prove extraordinarily difficult to dislodge. If only strong evidence is allowed in a reasonably large scientific community, however, information cascades will almost never result in false consensus. To demonstrate this, I developed an agent-based computer simulation. The simulation shows how allowing weak evidence in science leads to a risk of an information cascade to a false conclusion – a risk not present with strong evidence, or in the case of an individual reasoner. This favors SSE over PTE as the scientific evidentiary standard. We turn now to the details of the simulation.

4.1. *Overview*

The purpose of the model is to explore how the effect of evidence on scientists' beliefs is affected by the social structure of science. I designed the model in NetLogo, software designed for developing agent-based models. The model has only one kind of agent, representing researchers. Researchers consist of a CREDENCE[9] in the hypothesis in question, a BELIEF calculated from that credence, and a set of one-directional outgoing LINKS to other researchers, which represents the set of other scientists whose opinions they respect. Additionally, a global variable EVIDENCE-STRENGTH represents the strength of the kind of evidence available to all researchers, and another global variable HYPOTHESIS-IS-TRUE represents whether the hypothesis in question is true or not.

4.2 *Initialization and processes*

When the simulation is initialized, the set of researchers is created. Each consecutive researcher is assigned a set of LINKS to agents chosen randomly from the set of agents, weighted by the number of ingoing LINKS each agent already possesses plus one. For example, an agent with three in-LINKS is four times as likely to be selected as one with no in-LINKS. This yields a scale-free network,[10] which is the type of network exhibited by citation patterns in the sciences (de Solla Price 1965). Finally, researchers are randomly assigned an initial CREDENCE between 0.3 and 0.7.

9 Disclaimer: the names of simulation parameters should not be taken too seriously. BELIEF, for instance, should not be equated with 'belief', where 'belief' is our best philosophical account of belief. Likewise for CREDENCE, EVIDENCE-STRENGTH, etc. These parameters are merely highly idealized representations used to model epistemic agents.

10 The same simulation run on random networks produces similar results.

The model has two steps. First, each researcher individually assesses evidence. In each individual case, the valence of the evidence is determined probabilistically based on EVIDENCE-STRENGTH. For example, if EVIDENCE-STRENGTH is 0.6, each researcher has a 60% chance of finding evidence which agrees with HYPOTHESIS-IS-TRUE, and a 40% chance of finding misleading evidence. Each researcher then updates their CREDENCE on this evidence by Bayesian conditionalization. After updating, each researcher recalculates their BELIEF, believing the hypothesis only if their CREDENCE is greater than 0.5.

Second, each researcher observes the BELIEF of researchers it has outgoing LINKS to, and adjusts its CREDENCE on the basis of what it perceives the general attitude to be. Specifically, the researcher determines the percentage of peers who believe the hypothesis, then takes the midpoint between that percentage and its present CREDENCE to determine its new CREDENCE. For example, if an agent has a CREDENCE of 0.4 and 2 out of 10 peers believe the hypothesis, the agent will adopt a new CREDENCE of 0.3.

I have two comments on this second process, since this is where the effects of the social network come into play. First, that researchers only attend to their peers' beliefs and not their credences is justified because while we have at least rough first-person access to our own degrees of belief, we generally don't have precise third-person access to the degrees of belief of others. We have excellent third-person access, though, to the polar beliefs of others, and the model design reflects these facts.

Second, the formula I use to model how researchers take into account the beliefs of peers is admittedly simplistic. Starting simple, however, has been a fruitful approach for researchers modeling the social structure of science. Drawing on Kitcher's (1993) seminal work on formally modelling the social structure of science, a number of philosophers of science have used agent-based models to understand joint epistemic activity. The simplifications in my model resemble those in models prominent in the literature. There is precedent, for instance, in having agents assess how many of their peers have polar belief in a hypothesis (Zollman 2010). And Grim *et al.* (2011) model social influence on belief merely by having the agent average the credences of all its neighbors and adopt that average as its own credence. While even more simple than the analogous mechanism in my model, their formula captures enough of the phenomenon to make the result plausible, but is simple enough that it's clear why the result occurs. In making extraordinary simplifications, I am following the lead of what has been a fruitful methodology, as well as yielding to necessity. The psychological mechanisms of resolving near-peer disagreement are not yet well understood, so keeping the model simple allows us to aim for rough similarity with a broad range of plausible psychological mechanisms. In particular, the method I use has several features which are both realistic-looking and supported by experimental work on attitude change (Petty and Wegener 1998): agents take into account particularly the opinions of those they regard as reliable, agents weight their own opinion more heavily than that of any other individual, and agents care more about the general consensus than the testimony of individuals. So although the formula I use is simplistic, the results may still be illuminating.

4.3 Run parameters

For the purpose of analysis, I ran the simulation 60,000 times. This included runs with EVIDENCE-STRENGTH set at 0.55, 0.6, 0.65, 0.7, 0.75, and 0.8 for 1000 runs at

each value with HYPOTHESIS-IS-TRUE set to true, and an additional 1000 runs at each value with HYPOTHESIS-IS-TRUE set to false, for a total of 2000 runs at each value. This was repeated five times, with each researcher having outgoing LINKS to 5, 10, 15, 20, and 25 others in each respective repetition. Each run included 200 researchers.[11]

4.4 *Output and analysis*

Since the purpose of this simulation is to determine the interaction between evidence strength and the social nature of scientific reasoning, the primary output measure, PROPORTION-CORRECT, is the proportion of the research population with a true belief. For instance, if the hypothesis is false and 60 out of 200 agents believe it is false, then PROPORTION-CORRECT would be 0.3. For each run of the simulation, I had NetLogo report PROPORTION-CORRECT after setup (pregame score), after researchers update on the evidence (halftime score), and after researchers adjust their beliefs based on their social network (final score). This allows us to differentiate between the effects of initial evidence assessment and the resolution of peer disagreement, which is necessary because we are particularly interested in the latter.

The effect of initial evidence assessment is precisely what we would expect: There is some chance that evidence will mislead individuals when it is weak, and we see that here. 2711 out of 10,000 runs at evidence strength 0.55 saw a drop between pregame score and halftime score, indicating an increase in false belief across the population, as did 649 at strength 0.60 and 38 at strength 0.65. There was never an increase in false belief with stronger levels of evidence. This doesn't tell against PTE, since despite this risk of error with weak evidence the expected epistemic payoff is positive. I report these results only for purposes of comparison with the data of interest.

The real question is what happens when researchers pay attention to their social environment. Does resolving disagreement at low evidence strength lead to epistemically detrimental information cascades? In other words, do a significant number of runs increase in false belief between halftime and the final score? Yes: as shown in Figure 1.

At evidence strength 0.55, 4257 – well over a third of runs – show an increase in false belief when researchers take the beliefs of their peers into account. 2459 runs have the same result at evidence strength 0.60, and at higher levels of evidence strength the number of runs with a loss in true belief falls off sharply.

11 More than one reader has asked about whether the results reported below are robust to increasing the level at which a CREDENCE is treated as a BELIEF. In response, I ran the same set of simulations with BELIEF requiring a CREDENCE of at least 0.55, and again with the same parameter set at 0.60. The outcome differed from what is reported below in two respects. First, the number or runs where a population moves from true to false belief was lower at low levels of evidence strength, and tapered off more slowly as evidence strength increased (there were still no runs where this occurred at the highest levels of evidence strength). This difference was more pronounced at 0.60 than at 0.55. Second, the number of runs with accuracy loss due to social assessment increases significantly at 0.55, and even more so at 0.60. At low levels of evidence strength, it occurs the majority of time when the hypothesis in question is true. This appears to be because the more stringent criterion for belief leads to an appearance of less general confidence in the hypothesis. The effect, however, diminishes rapidly at higher levels of evidence strength. This is a different social epistemological benefit of excluding weak evidence than the one I draw on below, but either way excluding weak evidence has a notable quasi-qualitative effect.

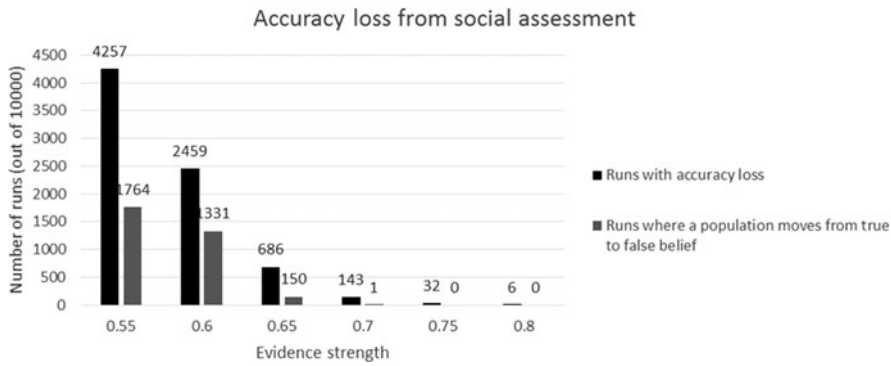

Figure 1. Summary of simulation results.

The key scenario, however, is when the social network itself creates a false consensus – when a population whose members mostly hold the true belief is converted into a population with a majority of members holding the false belief. At the lower levels of evidence strength, 0.55 and 0.6, this occurred in 1764 and 1331 out of 10,000 runs, respectively. At the highest levels of evidence strength, social assessment was never able to overturn a general acceptance of the truth. The opposite effect –a population with generally false belief converting to a population with generally true belief –is less frequent, occurring in 1296 runs at evidence strength 0.55, and 382 runs at evidence strength 0.6.

What does this mean? It doesn't mean, on its own, that we should ban evidence of strength 0.60[12] or lower from scientific practice, since despite the high rate of error it still might maximize expected epistemic payoff to use the weak evidence. But it does mean – and this is all I think the model reliably demonstrates – that weak evidence does frequently lead to quick, but inaccurate consensus. The simulation suggests that scenarios such as the moon-cheese fable will be more than rare oddities. Such scenarios are bad news for science not because they don't maximize expected epistemic accuracy in the short term, but because they put a halt to progress on questions that should remain open. One of the lessons of modeling the social structure of science has been that in the long run, quick consensus can be hard to displace and tends to be less accurate than the results of long-term debate (Zollman 2007; Grim *et al.* 2011). I'm making a similar, but distinct point. The short-term maximization of epistemic accuracy obtained by utilizing low quality evidence isn't worth it, I argue, because those cases where a quick false consensus emerges are too heavy a cost to pay. A false consensus generally impedes scientific progress more than leaving a question open.

The negative consequences of researchers accepting a hypothesis as accepted by consensus are severe. Not only does the field largely stop looking for what is in fact the right answer to that particular research question, but even worse, that false claim can become an accepted presupposition for future theorizing and thus skew the interpretation of future

12 I don't think we should attach any special significance to these numbers in particular. It would be a mistake, for example, to reformulate the absolute criterion in SSE to specifically draw the line at 60% reliability. The mistake would lie not only in taking the details of this particular model too seriously, but also in forgetting that SSE is the result of multiple factors, not just the one the model is designed to identify.

results. This latter effect particularly gives us good reason to want to prevent any significant number of false consensuses in science. What we learn from the simulation is that false consensuses as a result of the social structure of science are a problem mostly when our primary sources of evidence are weak, so this gives us a reason to favor a standard of evidence which forbids the use of weak evidence. In other words, it gives us yet one more reason to prefer SSE to PTE.

5. MOVING PARTS

I've argued that SSE is superior to PTE on three different fronts, but in each case the possibility remains that we could address the shortcoming of PTE by adjusting some feature of the structure of science besides the standard of evidence. In this section, I'll briefly explain why these alternatives are inferior to using SSE as the remedy.[13]

We could address the problem of perverse incentive, for example, by changing the incentive structure of science rather than changing the standard of evidence. This could go two ways. First, we could radically alter how scientists get credit. Putting aside whether this is a realistic possibility, it would be a bad idea, given how valuable credit-seeking along extant lines is for achieving epistemic ends (Stephan 1996; Bruner 2013; Heesen, Forthcoming). SSE allows us to reap the benefits of the extant incentive structure while mitigating some of the costs. Second, we could just give scientists less credit for publishing work that uses weaker evidence. This approach, however, would effectively be the implementation of a norm similar enough to SSE that I don't think it's worth quibbling about the difference.

As for the problem of cognitive bias, we could look for other means of mitigating the effects of bias on the production of science besides implementing SSE. We certainly should, since SSE doesn't address all forms of bias, nor does it completely mitigate the problems that it does address. But unless there is another means of mitigation that completely obviates the benefits of SSE, enforcing SSE remains an important part of the structure of science.

The problem of harmful information cascades might be addressed by reducing information flow between scientists as an alternative to SSE. But this is likely to do more harm than good, since lack of information flow causes its own problems, such as publication bias (Rosenthal 1979; Kicinski 2014). SSE is a means to address the problem which has the additional positive effect of dealing with the other epistemic issues discussed above, so it is preferable to reducing information flow.

6. FURTHER DISCUSSION

We began with a puzzling fact. Not everything which would be confirmatory to an ideally rational being is accepted as confirmatory in science. I presented this puzzle as a competition between two norms: the principle of total evidence (PTE), which states that rational confirmation uses all available evidence, and the scientific standard of evidence (SSE), which states that scientific confirmation uses sources of evidence which are among the best available.

I've attempted to give a partial answer as to why the scientific community is rationally justified in adhering to SSE rather than PTE. I first argued that by forbidding the use of

13 The need for this section was raised by a helpful reviewer.

weaker sorts of evidence we better incentivize individual scientists to make optimal contributions to joint epistemic projects. In support of this reasoning, I presented two case studies of contemporary sciences where leading figures have identified major problems in their fields resulting from too lax a standard of evidence. I then argued that PTE is only optimal for ideal agents, and gave examples from experimental psychology of systematic biases in human reasoning which entail that SSE yields better epistemic returns for actual human cognitive agents. Finally, I argued that because confirmation in science is inherently social, weaker types of evidence can lead to harmful false consensus.

I make no claim that these explanations are exhaustive, but I do think that they account for a significant portion of why something like SSE should be accepted in most scientific disciplines. Some readers may worry that it isn't true that something like SSE is accepted across the sciences. It doesn't take a professional sociologist to notice that as a matter of fact, however much scientists might profess adherence to something like SSE, scientific theorizing and argumentation frequently does involve invocation of weaker sorts of evidence such as intuition and anecdote. I have two responses to this objection. First, even deeply rooted norms are not observed perfectly. Even though modern science generally does accept something like SSE, we should still expect to find violations. But we should also find that those violations, if noticed, are policed from within the discipline. This is precisely what's going on in the case studies from generative linguistics and social psychology. Second, SSE is only about what is permissible for use in confirmation. Weaker sorts of evidence may still play important roles in the context of scientific discovery, such as in helping scientists formulate hypotheses, design experiments, and create tentative theories. Given that discovery is as much a part of science as is justification, we should expect to see things like anecdotes and intuition-data appearing in scientific practice.

I'll close by suggesting some implications of this account of why not all evidence is scientific evidence. First, while SSE is not a solution to the problem of demarcation between science and pseudoscience, it does provide a heuristic that can help with some of the questions an account of demarcation is meant to answer. For example, in scientific settings it justifies us in ignoring claims of alternative medicines if they are based entirely on weaker sorts of evidence. In fact, SSE indicates that it would be wrong to take seriously those claims or the evidence supporting them.

A second implication of my arguments is that other epistemic contexts may require a stricter standard of evidence. The obvious example is the courtroom, where in many nations certain types of evidence are already excluded. My treatment of the scientific standard of evidence in terms of purely epistemic reasons for adopting it could be adapted to the legal context, though non-epistemic values may also need to be accounted for. We could also give similar treatments to other institutional and organizational epistemic contexts, such as policy-making, intelligence-gathering, and market research. Furthermore, I think some of the reasons I've outlined here apply even to individual contexts, and while I won't argue for it here, there's an argument to be made for why not all evidence is evidence for individual human beings.[14] I'm content for the time being, however, to have shown merely how scientific reasoning benefits from its higher standard of evidence.

14 Kadane *et al.* (2008) provide formal arguments that this is the case in certain situations for Bayesian agents. I suspect that we can find more and better arguments if we look at much less-idealized reasoners, as I have done in this paper.

REFERENCES

Achinstein, P. 1995. 'Are Empirical Evidence Claims A Priori?' *British Journal for the Philosophy of Science*, 46: 447–73.

—— 2001. *The Book of Evidence*. Oxford: Oxford University Press.

Banerjee, A. V. 1992. 'A Simple Model of Herd Behavior.' *Quarterly Journal of Economics*, 107: 797–817.

Baumeister, R. F., Vohs, K. D. and Funder, D. C. 2007. 'Psychology as the Science of Self-reports and Finger Movements: Whatever Happened to Actual Behavior?' *Perspectives on Psychological Science*, 2: 396–403.

Bikhchandani, S., Hirshleifer, D. and Welch, I. 1998. 'Learning From the Behavior of Others: Conformity, Fads, and Informational Cascades.' *Journal of Economic Perspectives*, 12: 151–70.

Bruner, J. P. 2013. 'Policing Epistemic Communities.' *Episteme*, 10: 403–16.

Carnap, R. 1947. 'On the Application of Inductive Logic.' *Philosophy and Phenomenological Research*, 8: 133–48.

—— 1962. *Logical Foundations of Probability*, 2nd edition. Chicago, IL: University of Chicago Press.

de Solla Price, D. J. 1965. 'Networks of Scientific Papers.' *Science*, 149(3683): 510–15.

Gigerenzer, G. and Brighton, H. 2009. 'Homo Heuristicus: Why Biased Minds Make Better Inferences.' *Topics in Cognitive Science*, 1: 107–43.

Good, I. J. 1967. 'On the Principle of Total Evidence.' *British Journal for the Philosophy of Science*, 17: 319–21.

Grim, P., Singer, D. J., Reade, C. and Fisher, S. 2011. 'Information Dynamics Across Sub-Networks: Germs, Genes, and Memes.' In *AAAI Fall Symposium: Complex Adaptive Systems*.

Heesen, R. Forthcoming. 'Communism and the Incentive to Share in Science.' *Philosophy of Science*.

Ioannidis, J. P. 2005. 'Why Most Published Research Findings are False.' *PLoS Medicine*, 2(8): e124.

Kadane, J. B., Schervish, M. and Seidenfeld, T. 2008. 'Is Ignorance Bliss?' *Journal of Philosophy*, 105 (1): 5–36.

Kahneman, D., Krueger, A. B., Schkade, D., Schwarz, N. and Stone, A. A. 2006. 'Would you be Happier if you were Richer? A Focusing Illusion.' *Science*, 312(5782): 1908–10.

Kicinski, M. 2014. 'How Does Under-Reporting of Negative and Inconclusive Results Affect the False-Positive Rate in Meta-Analysis? A Simulation Study.' *BMJ Open*, 4(8): e004831.

Kitcher, P. 1993. *The Advancement of Science: Science Without Legend, Objectivity Without Illusions*. Oxford: Oxford University Press.

Nickerson, R. S. 1998. 'Confirmation Bias: A Ubiquitous Phenomenon in many Guises.' *Review of General Psychology*, 2: 175–220.

Nosek, B. A., Spies, J. R. and Motyl, M. 2012. 'Scientific Utopia II. Restructuring Incentives And Practices To Promote Truth Over Publishability.' *Perspectives on Psychological Science*, 7: 615–31.

Petty, R. E. and Wegener, D. T. 1998. 'Attitude Change: Multiple Roles for Persuasion Variables.' In D. Gilbert, S. Fiske and G. Lindzey (eds), *The Handbook of Social Psychology*, 4th edition, Vol. 1, pp. 323–90. New York, NY: McGraw-Hill.

Pfungst, O. 1911. *Clever Hans (The Horse of Mr. von Osten): A Contribution To Experimental Animal And Human Psychology*. New York, NY: Holt, Rinehart and Winston.

Rosenthal, R. 1979. 'The File Drawer Problem and Tolerance for Null Results.' *Psychological Bulletin*, 86: 638–41.

Simmons, J. P., LeBoeuf, R. A. and Nelson, L. D. 2010. 'The Effect of Accuracy Motivation on Anchoring and Adjustment: Do People Adjust from Provided Anchors?' *Journal of Personality and Social Psychology*, 99: 917–32.

——, Nelson, L. D. and Simonsohn, U. 2011. 'False-positive Psychology Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant.' *Psychological Science*, 22: 1359–66.

Simonsohn, U., Nelson, L. D. and Simmons, J. P. 2014. 'P-curve: A Key to the File-drawer.' *Journal of Experimental Psychology: General*, 143: 534–47.

Stephan, P. E. 1996. 'The Economics of Science.' *Journal of Economic Literature*, 34: 1199–235.

Strack, F., Martin, L. L. and Schwarz, N. 1988. 'Priming and Communication: Social Determinants of Information Use in Judgments of Life Satisfaction.' *European Journal of Social Psychology*, 18: 429–42.

Tversky, A. and Kahneman, D. 1974. 'Judgment under Uncertainty: Heuristics and Biases.' *Science*, 185(4157): 1124–31.

Wasow, T. and Arnold, J. 2005. 'Intuitions in Linguistic Argumentation.' *Lingua*, 115: 1481–96.

Wilson, T. D., Houston, C. E., Etling, K. M. and Brekke, N. 1996. 'A New Look at Anchoring Effects: Basic Anchoring and its Antecedents.' *Journal of Experimental Psychology: General*, 125: 387–402.

Zollman, K. J. 2007. 'The Communication Structure of Epistemic Communities.' *Philosophy of Science*, 74: 574–87.

—— 2010. 'Social Structure and the Effects of Conformity.' *Synthese*, 172: 317–40.

CARLOS SANTANA is based at the Department of Philosophy, University of Utah. His research interests are Philosophy of Language Science and Philosophy of Environmental Science.