Bias neglect:
A blind spot in the evaluation of scientific results

Brent Strickland

Yale University

brent.strickland@yale.edu

Hugo Mercier

CNRS – L2C2

hmercier@isc.cnrs.fr

Running Head : Perception of Bias
Address for correspondence : Brent Strickland
Yale University Dept. of Psychology
Box 208205
New Haven, CT 06520
Email : brent.strickland@yale.edu
Phone/Fax : 203-432-6451
Word Count : 5681 (main text, author note)

Abstract (146 words)

Experimenter bias occurs when scientists' hypotheses influence their results, even if involuntarily. Meta-analyses (e.g. Rosenthal & Rubin, 1978) have suggested that in some domains, such as psychology, up to 1/3 of the studies could be unreliable due to such biases. A series of experiments demonstrates that while people are aware of the possibility that scientists can be more biased when the conclusions of their experiments fit their initial hypotheses, they robustly fail to appreciate that they should also be more skeptical of such results. This is true even when participants read descriptions of studies that have been shown to be biased. Moreover, participants take other sources of bias—such as financial incentives—into account, showing that this bias neglect may be specific to theory driven hypothesis testing. In combination with a common style of scientific reporting, bias neglect could lead the public to accept premature conclusions.

Keywords
Experimenter bias, scientific reasoning, decision making, epistemic vigilance

The media can bring scientific results to a very wide audience. Unfortunately, a substantial number of these results turn out to be of dubious value (Gonon et al, 2012; Ioannidis, 2005) and they can cause significant damage—as when a link between vaccines and autism was suggested (RETRACTED: Wakefield et al., 1998). While it is critical that the public be kept informed of scientific developments, it should not accept scientific reports uncritically (Kahan, 2010). Below we describe the discovery of a blind spot in the public's naïve understanding of science that could have important implications for the public's understanding of science.

Scientists' mental states—their hypotheses, expectations and beliefs—sometimes influence their results in a phenomenon referred to as *experimenter bias*. In typical demonstrations, experimenters are asked to test either hypothesis A or hypothesis B, and are shown to unintentionally bias their results so that they fit with their given hypothesis (Rosenthal & Lawson, 1964; Yank, Rennie & Bero, 2007; Strickland & Suben, in press). This tendency to bring about desired but unreliable results is surprisingly robust. Rosenthal & Rubin (1978) carried out a meta-analysis of 345 studies, each designed to explicitly examine experimenter bias. They showed that across a range of topics in psychology, more than a third yielded results that were unreliable due to experimenter bias. The average effect size of bias was also surprisingly large (Cohen's d=.7).

Outside the context of experiments specifically designed to test for the presence of experimenter bias, recent meta-analyses of reported scientific findings in the literature have also demonstrated a tendency for researchers to find what they are looking for. For instance, one recent meta-analysis by

Wilgenburg and Elgar (2013) analyzed the results of 76 studies of nestmate recognition in ants. The baseline theoretical expectation by researchers in this community is that there should be less aggression between ants that share a nest than between ants that do not share a nest. Upon comparing the results of those studies in which researchers were blind to experimental condition (when coding ant behaviors as being aggressive or non-aggressive) with results from studies in which researchers were aware of experimental condition, they found that blind studies were more likely to report results which went counter to the standard scientific consensus than those studies in which researchers were not blind to condition. Blind studies were much more likely to report aggression between nestmates (73%) than non-blind studies (21%), and the average effect between nestmate vs. non-nestmate means was robustly lower for blind studies (d=1.38) than for non-blind studies (d=2.76). These findings, as well as the Rosenthal meta-analysis, strongly suggest that experimenter bias is a robust phenomenon that can impact the credibility of scientific reports.

Here we ask whether people understand that they should be more skeptical of results that fit with experimenter hypotheses, given the known possibility of bias. On the one hand, they could put too much stock in the phenomenon, using it to dismiss a result simply because they disagree with the scientists' views. On the other hand, they could also fail to pay sufficient attention to scientists' beliefs and completely ignore the possibility of experimenter bias.

Anecdotal evidence suggests that people will take scientists' mental states into account when evaluating their conclusions since scientists' beliefs are

sometimes used to attack their results. Thus, for example, both sides of the climate change debate have been accused of modifying research results to support political agendas (McKitrick, 2011; Michaels, 2008).

Moreover, theoretical work surrounding the nature and development of cynicism and epistemic vigilance suggests that the inferred mental states should matter for the evaluation of scientists' conclusions (Eagly, Wood, & Chaiken, 1978; Mills & Jellison, 1967; Heyman, Fu, & Lee, 2007; Mills & Keil, 2005). By and large, this literature has demonstrated that both children and adults are sensitive to information about other people's knowledge, intentions, and desires, and are capable of using this information for the purposes of evaluating their claims. For example, from around four years of age, young children can use mental state attributions to gauge speakers' competence and to evaluate their statements on that basis—for instance, young children understand that lack of relevant knowledge precludes some speakers from generating informed statements (Baron-Cohen, Leslie, & Frith, 1985; Sodian, 1988; Wimmer & Perner, 1983). More relevantly to the current experiments, existing empirical work has also shown that both adults (e.g., Eagly, Wood, & Chaiken, 1978; Mills & Jellison, 1967) and children (Heyman, Fu, & Lee, 2007; Mills & Keil, 2005) can also use speakers' desires to evaluate their statements. In particular, children and adults discount self-interested statements (i.e. statements that fulfill the speaker's desires), and a conclusion that fits with one's hypothesis might be considered as a self-interested statement, given a general preference for being right.

These experiments, however, did not bear directly on scientific communication, which might be evaluated differently from other types of

communication. Indeed, science may be an important exception to the common belief found in the literature on skepticism that mental states are central to evaluating claims. Sociologists of science point out that 'scientific facts' tend to acquire an objective character, ignoring the role played by the scientific community and its potential biases (e.g, Collins, 1983; Fleck, 1981). Moreover, people do not discount biased advice from expert advisors as much as they should, even when that bias is explicitly disclosed (Cain, Lowenstein & Moore, 2005). Thus scientists might be seen as producing objective facts, their beliefs having no impact on their conclusions.

In order to better understand how mental state information influences people's evaluations of scientific claims, a series of experiments, we ask the following two questions: Do people attribute bias to scientists based on the scientists' expectations? If yes, do people take this bias into account when evaluating the scientists' conclusions?

Study 1

In order to test the influence of researcher expectations on the evaluation of scientific results, participants read about a fictional experiment describing scientists who either succeeded or failed in obtaining a desired result.

Methods

Participants

157 participants were recruited through the Amazon Mechanical Turk website. They were paid $0.1 for their participation, a normal rate for this type of

task in Mechanical Turk. All participants had to be in the US at the time of the experiment. Several published studies already rely on this sample (e.g. DeScioli & Kurzban, 2009; Mercier & Strickland, 2012), and specifically designed experiments have established its reliability (e.g. Buhrmester, Kwang, & Gosling, 2011; Paolacci, Chandler, & Ipeirotis, 2010). Tables 1 to 3 provide an overview of the demographics for all the experiments, participants having been recruited through the same mean.

INSERT TABLES 1-3

Design

We employed a between participant, 2 x 2 design varying the intention of the scientists (verification vs. falsification of their original hypothesis) and their success in fulfilling their intentions (success vs. failure). Participants read one of four variations on the following story:

> *Two groups of researchers in England were debating how kidney hormone production is controlled in the body. The kidney produces a hormone called Calcitriol. The researchers from the north of England thought that a chemical called AXF1, which is produced in the thyroid, 'turns on' the part of the kidney that is responsible for making Calcitriol. The researchers from the south of England instead believed that a chemical called CVR2, which is produced in the hypothalamus, is responsible for turning on Calcitriol production in the kidney. Both teams agree that it's either one of the two chemicals that is responsible for the production of Calcitriol, and that it cannot be the two of them together.*

The researchers from the south decided to try to confirm their theory. In order to do this they artificially manufactured the CVR2 compound [the one they believe responsible for hormone production] in pill form and simply had one group of test subjects ingest it. Another group of subjects instead took a placebo pill.

Just as they expected, the researchers from the south found that subjects who had taken CVR2 produced more Calcitriol than subjects who had taken the placebo drug. Thus they concluded that their initial theory was right and that it is CVR2, and not AXF1, that is responsible for the production of Calcitriol.

In order to create the Falsification/Failure condition, these paragraphs were altered as follows:

The researchers from the north decided to try to falsify the southerners' theory. In order to do this they artificially manufactured the CVR2 compound in pill form and simply had one group of test subjects ingest it. Another group of subjects instead took a placebo pill.

To their surprise, the researchers from the north found that subjects who had taken CVR2 produced more Calcitriol than subjects who had taken the placebo drug. Thus they concluded that their initial theory was wrong and that it is CVR2, and not AXF1, that is responsible for the production of Calcitriol.

The Verification/Failure and Falsification/Success were created from the relevant mix of the first and second paragraph, with the results adjusted accordingly.

Participants were asked the following three questions (in this order[1]) and answered them on a 1 (not at all) to 7 (very much) scale:

Do you think that the southerners [northerners] were biased in drawing their conclusion? [Bias question]

How much do you think that the experiment of the southerners [northerners] supports their conclusion? [Support question]

How much do believe the conclusion of the southerners [northerners]? [Belief question]

Results and discussion

A 2 (failure vs. success) x 2 (verification vs. falsification) between subjects ANOVA revealed no significant main effects of verification vs. falsification on any of the dependent variables in Studies 1, 2, or 3. Therefore we report only the results related to researcher success vs. failure in all studies.

Participants believed that scientists were more biased when they found the result that suited their hypothesis (Bias question: 4.40 vs. 2.54, $F(1,153)=42.330$, $p<.001$, $\eta_p^2=.22$). This perceived bias did not influence either the degree to which participants believed the experiment supported the conclusion (Support question 4.41 vs. 4.58, $F(1,153)=.409$, $p=.52$, $\eta_p^2=.00$), nor how much

[1] We asked the "bias" question first in order to make the perception of bias salient, and thus to give participants every chance to use this information. We are of course aware that this introduces the possibility of order effects and therefore also counterbalance the order of questions in subsequent studies below.

they believed the conclusion itself (Belief question 4.37 vs. 4.7, $F(1,153)=1.469$, $p=.23$, $\eta_p^2=.01$).

These results suggest people recognize the potential biasing effect of scientists' beliefs, but do not take this into account when evaluating scientists' conclusion nor when considering whether the experiment supports the conclusion.

Study 2

In Study 1, participants were provided with a full description of the experiment conducted by the scientists and the logic of the experiment may have taken precedence of consideration of bias. In order to test this, Study 2 had participants read an abstract version of the fictional Experiment from Study 1, which should prevent participants from relying on their own evaluation of the experiment to evaluate its conclusion.

Methods

Participants

171 participants took part in Study 2.

Design

Study 2 was identical to Study 1 except that the description of the experiment was underspecified. More specifically, the last two paragraphs of the text of Experiment 1a were modified to trim any detail about the study (here the Verification/Success condition):

The researchers from the south decided to try to confirm their theory. They performed an experiment involving chemical compounds given to different groups of participants.

The researchers got the results they expected. They concluded that their initial theory was right and that it is CVR2, and not AXF1, that is responsible for the production of Calcitriol.

Results and discussion

Participants attributed more bias to the scientists whose conclusion matched their original hypothesis (Bias question: 4.46 vs. 2.61, $F(1,167)=53.82$, $p<.001$, $\eta_p^2=.24$). However, in contrast to Study 1, this perceived bias did influence belief in the conclusion: when the scientists found the result they were looking for, participants were less likely to believe their conclusion (4.37 vs. 4.7, $F(1,166)=51.84$, $p<.001$, $\eta_p^2=.11$). There was no significant difference for the Support question (4.79 vs. 4.790, $F(1,167)=.011$, $p=.92$, $\eta_p^2=.00$).

Study 2 shows that people can not only attribute bias to scientists based on the fit between their beliefs and their conclusions, but also that they take this bias into account absent other ways to evaluate the conclusion. This finding additionally suggests that people think of the degree of support that the data lend to a conclusion is less influenceable by experimenter expectation than the quality of the data itself.

Study 3

In Study 1, participants presumably relied on their assessment of the experiment described to evaluate its conclusion, whereas in Study 2, in the absence of a sufficient description, they relied on the perceived bias of the experimenters. In neither study was there any obvious flaw in the experiment described to the participants. If such a flaw were introduced, participants could rely on their assessment of the experiment to evaluate its conclusion, as in Study 1. However, they might also take perceived bias into account in order to gauge the importance of the flaw—for instance, is it a mere mistake or an intentional oversight?

Methods

Participants

163 participants took part in Study 3.

Design

Study 3 is identical to Study 1 with a blatant flaw added in the description of the experiment. One sentence was modified from Study 1, the same in all conditions. Instead of specifying, "Another group of subjects instead took a placebo pill," it said, "They did not use a control group taking a placebo pill."

Results and discussion

As in both previous studies, participants believed that the research groups were more biased when they found the desired results (Bias question 5.76 vs. 3.44, $F(1,159)=65.90$, $p<.001$, $\eta_p^2=.29$). As in Study 2, participants took potential

bias into account in answering the Belief question (2.87 vs. 3.79, $F(1,159)=10.73$, $p<.001$, $\eta_p^2=.06$) but not the Support question (3.18 vs. 3.62, $F(1,159)=2.35$, $p=.13$, $\eta_p^2=0.02$).

These findings suggest that when people are confronted with poor evidence for a scientific conclusion, they turn to other cues to evaluate the conclusion such as the potential bias introduced by the scientists' beliefs. Moreover, the results from the current study replicate the finding from Study 2 whereby people think the quality of the data is more influenceable by experimenter expectation than the degree of support that the data lend to a conclusion.

Studies 4 and 5

In Studies 1-3, participants read about a fictional physiology experiment. Here we replicate these findings using a different scientific context: neuroscience. Additionally, given that Studies 1-3 established that participant evaluations of the support between experiment and conclusion is robustly insensitive to experimenter belief, even in extreme circumstances like that of Studies 2 and 3, we concentrate more directly on the relationship between the appreciation of potential bias and the willingness to discount one's belief in scientists' conclusions.

Methods

Participants

90 participants took part in Study 4, 91 in Study 5.

Design

Studies 4 and 5 are similar to Studies 1 and 3 in that Study 4 gave a full description of the experiment while Study 5 described that same experiment with an embedded methodological flaw. There were three additional differences. First, the topic was different (physiology vs. neuroscience). Second, only the success and failure conditions were maintained. Third, two questions were used: the Bias question and the Belief question, in counterbalanced order.

Here is a sample text used in Study 4, Failure condition:

Two groups of researchers were debating how the brain represents "simple physics" like the fact that two solid objects cannot pass through one another. Researchers from the University of Northumbria thought that the medial temporal cortex ("MT") was responsible for processing naive physics while researchers from Lancaster University instead believed that it was the intraparietal sulcus ("IPS"). Both teams agreed that it was either one of two brain regions that was responsible for naive physics, and that it could not be the two of them together.

The researchers from Lancaster University decided to try to confirm their theory. In order to do this, they performed a type of experiment called "Transcranial Magnetic Stimulation" (TMS for short) which allows researchers to temporarily de-activate very specific brain regions but leave the other brain regions fully functional.

> *To their surprise, the researchers from Lancaster University found the opposite results of what their theory predicts. Thus they concluded that their initial theory was wrong, and that it is MT, and not IPS, that is responsible for naive physics.*

In the success condition, the last paragraph was altered as follows:

> *Just as they expected, the researchers from Lancaster University found the results predicted by their theory. Thus they concluded that their initial theory was right, and that it is IPS, and not MT, that is responsible for naive physics.*

For Study 5 the same texts were used, with the following sentence added at the end of the second paragraph: *"They did not use a control group in their experiment."*

Results and discussion

The results replicate those of Studies 1 and 3. In Study 4, planned comparisons revealed that participants rated the successful scientists as being more biased than the unsuccessful ones (4.53 vs. 2.98, $t(88)=5.02$, $p<.001$, $\eta_p^2=0.22$), but they did not take the scientists' beliefs into account when evaluating the conclusion (4.36 vs. 4.64, $t(88)=.9$, $p=.37$, $\eta_p^2=.00$). In Study 5 participants did not only attribute more bias to the scientists whose results fit with their hypotheses (4.90 vs. 3.51, $t(89)=3.61$, $p<.01$, $\eta_p^2=.13$), but they also believed less in their conclusion (3.10 vs. 4.07, $t(89)=2.94$, $p<.01$, $\eta_p^2=.09$).

Study 6

In Studies 1 and 4, participants attributed bias to the scientists but did not take this bias into account when evaluating their conclusions. Since the stimuli

were fictional experiments, it is impossible to tell whether the participants *should* have taken perceived experimenter bias into account. Study 6 addresses this concern by describing to participants a study in which a robust effect of experimenter bias has been demonstrated.

Rosenthal and Lawson (1964) told one group of research assistants that they were testing "maze bright" rats that had been bred to perform very well on a maze-learning task. Another group was given "maze dull" rats, supposedly bred to perform poorly on such tasks. Each group showed that their rats performed in line to their supposed breeding, despite the fact that the two groups of rats actually did not differ in their breeding. Thus the scientists' beliefs altered the outcome of the experiment. Study 6 replicates Studies 1 and 4 while describing the Rosenthal and Lawson experiment to the participants.

It is worth noting that in Rosenthal and Rubin's 1978 review, animal learning studies were shown to be the most likely to produce such experimenter effects, with 73% showing an effect of bias. This type of study also had the highest average effect size with a Cohen's d of 1.73. Thus if participants are basing their judgments on the realities of science, in this case they should be particularly prone to discrediting the conclusions of a biased scientist.

Methods

Participants

60 participants took part in Study 6.

Design

Study 6 is similar to Studies 1 and 4, but with a description of the experiment of Rosenthal and Lawson (1964), here in the Failure condition:

> *In a laboratory, assistants were asked to perform an experiment. They were given two groups of rats and told to compare their performance on a maze task—finding a treat in a maze. The first group of rats was called "maze bright," and assistants were told that they had been bred to perform very well on maze tasks. The second group of rats was called "maze dull," and assistants were told that they have been bred to perform very poorly on maze tasks. The assistants hypothesized that because of their differing genetic make-ups, the rats in the "maze bright" group would more quickly learn to find their way around a maze than the "maze dull" group.*
>
> *The experiment consisted in launching the rats, one at a time, from the same point of the maze, and measuring how long it takes them to reach a treat that is always at the same place in the maze. For each rat, the experiment was performed once a day for five days in a row.*
>
> *The assistants observed that the "maze bright" rats did not learn to find the treats faster than the "maze dull" rats after several days of training. They thus concluded that their original hypothesis was wrong, and that the breeding techniques had not influenced rat learning ability.*

In the success condition, the last paragraph was altered as follows:

> *The assistants observed that the "maze bright" rats learned to find the treats significantly faster than the "maze dull" rats after several days of training. They*

thus concluded that their original hypothesis was correct, and that the breeding techniques influenced rat learning ability.

As in Studies 4 and 5, participants were asked about the bias of the experimenters and their belief in the conclusion, in counterbalanced order.

Results and discussion

Participants attributed more bias to the scientists who obtained results in line with their hypotheses (5.19 vs. 3.45, $t(58)=3.69$, $p<.001$, $\eta\rho^2=.19$), but they again failed to take this into consideration when evaluating the scientists' conclusion (4.39 vs. 4.76, $t(58)=.892$, $p=.38$, $\eta\rho^2=.01$). This result shows that participants fail to take bias into account even when they should.

Study 7

Studies 1, 4 and 6 have shown that although people think scientists can be biased by their beliefs, they assign bias a relatively low priority when evaluating the scientists' conclusions. Study 7 asks about the underlying mechanism behind this effect, and the discrepancy with other well-known effects showing that people do heavily discount their beliefs in biased testimony (e.g. Mills & Keil, 2005). One possible explanation is that the identity of the biased person is the driving factor behind bias neglect. Perhaps, for example, people reason that bias has less influence on scientists than on other people. In Study 7, we explicitly test this by having participants read fictional experiments involving either a scientist or a farmer.

Methods

Participants

134 participants took part in Study 7.

Design

In a 2 x 2 design, the first variable was the person forming the hypothesis being tested—a scientist or a farmer—and the second the outcome—success or failure. Participants read a fictional text about testing the efficiency of different types of manure on crop growth, here in the Farmer Failure condition:

> *Different types of manure can be used to fertilize crops. A farmer in Wisconsin did some extensive background reading on this topic, and came to strongly believe that horse manure was a more effective fertilizer than cow manure because of differences between the diets of horses and cows. In order to test his idea, he decided to do a test in which he fertilized three of his fields with horse manure and he fertilized another three fields with cow manure. Afterwards, he looked at the overall crop growth in each of the fields. In contrast with his original guess, he found that those fields that had been fertilized with cow manure produced more crops than those fields that had been fertilized with horse manure. Thus he concluded that his idea was wrong, and that cow manure is a more effective fertilizer than horse manure.*

In the Success condition, the last paragraph was altered as follows:

> *Consistent with his original guess, he found that those fields that had been fertilized with horse manure produced more crops than those fields that had been*

fertilized with cow manure. Thus he concluded that his idea was correct, and that horse manure is a more effective fertilizer than cow manure.

In the Scientist conditions, an agricultural scientist conducts an experiment (instead of doing a test) in order to test his theory (rather than his idea).

Results and discussion

Study 7 replicated the basic effect, and revealed that the profession of the actor (i.e. scientist vs. farmer) made no difference. A 2x2 ANOVA revealed that while people thought that the person whose results fit his hypothesis was more biased (3.91 vs. 2.15, $F(1,130)=37.99$, $p<.001$, $\eta_p^2=.23$), they did not discount that person's conclusion (4.69 vs. 4.99, $F(1,130)=1.49$, $p=.22$, $\eta_p^2=.01$). This pattern of results was not reliably different in the Scientist and Farmer conditions as evidenced by a lack of significant interaction between the profession and outcome on both the Bias question ($p=.21$, $\eta_p^2=.01$) and Belief question ($p=.33$, $\eta_p^2=.01$).

These results show that the reluctance to discount bias in forming one's conclusions is not connected to the status as scientist of the person providing the testimony since the neglect is equally strong for scientists and farmers. Thus some other factor must explain the difference between the current results and the more general tendency to discount one's own belief in biased testimony.

Study 8

Another possibility is that bias based on theories might be perceived as less likely to influence the outcome of an experiment than bias emanating from

sources reflecting other sources of self-interest, such as financial incentives. Thus bias neglect might be very specific not to the identity of the person involved but instead the type of activity she is engaging in, such as theory driven hypothesis testing. In order to test this, participants evaluated an experiment that was motivated either by a theory or by financial incentives.

Methods

Participants

68 participants took part in Study 8.

Design

The design was similar to the Scientist condition of Study 7, expect that now the scientist has a financial motivation. The text for the Failure condition read as follows (the changes for the Success condition were identical to those made in Study 8):

> *Different types of manure can be used to fertilize crops. An agricultural scientist in Wisconsin was approached by a group of horse owners who wanted to sell their manure. They paid the scientist to run an experiment to prove that horse manure is a more effective fertilizer than cow manure. In order to test the horse owners' theory, he decided to run an experiment in which he fertilized three local fields with horse manure and he fertilized another three fields with cow manure. Afterwards, he looked at the overall crop growth in each of the fields. In contrast with the horse owners' view, he found that those fields that had been fertilized with cow manure produced more crops than those fields that had been fertilized*

with horse manure. The scientist also concluded that the horse owners were wrong, and that cow manure is a more effective fertilizer than horse manure.

Results and discussion

Planned comparisons confirmed our original prediction. The scientist whose conclusion suited her financial backing was perceived as more biased (3.51 vs. 2.20, $t(66)=3.44$, $p<.01$, $\eta_p^2=.15$) and her conclusion was believed less (4.52 vs. 5.23, $t(65)=2.05$, $p<.05$, $\eta_p^2=.06$).

Participants thus take bias into account when evaluating a scientist's conclusion when that bias is financial rather than based on the scientist's hypotheses—although it should be noted that belief in the scientists' conclusion remains high, even with the conflicting financial incentives. This suggests that the reluctance to discount one's belief in biased scientific conclusions may be highly specific to the type of activity being carried out, and explains why the present findings differ from previous findings.

Discussion

The present studies demonstrate a potential blind spot in the evaluation of scientific results. Upon learning about a scientific experiment, people attribute more bias to the scientist when her conclusion fits with her initial hypothesis than when it does not. However, they fail to take this perceived bias into account when evaluating that same conclusion. This result is robust to variations in context (studies 1, 4) and does not seem to be influenced by the expertise of the person whose conclusion is being evaluated (study 7). Importantly, "bias

neglect" is also obtained when participants evaluate the conclusion of a study known to have been influenced by such bias (study 6). Together, these four studies show that people neglect a potentially significant source of bias.

An explanation for this 'bias neglect' is that people rely on other cues to evaluate the conclusion, chiefly, their own understanding of the experiment described. When the logic of the experiment is unclear (study 2) or flawed (study 3), participants take bias into account in their evaluation of the conclusion. People also take other sources of bias—such as financial motivations—into account when evaluating scientific conclusions (study 8), suggesting that 'bias neglect' is specific to theory driven hypothesis testing.

Theoretical implications

The literature on epistemic vigilance and skepticism has stressed that people are quite sensitive to speaker motivations in the evaluation of their claims. One example comes from a Mills and Keil (2005) study in which children were told a story about a character who would win a prize if certain conditions were met, but in the story it was ambiguous whether these conditions were met or not. Here is a verbatim description of one story from that study: "Michael was in a running race, and he and another boy finished the race close together (thus leaving it ambiguous who actually won). For the with-self-interest stories, the main character affirmed that the conditions for him or her to win the prize had been met; for the against-self-interest stories, the character denied that the conditions had been met and claimed that he or she should not win the prize. It was left ambiguous what the main character actually knew about the outcome."

This study showed that by second grade, children were sensitive to self-interest in their evaluation of the character's statement, and thus believed the character in the against self-interest stories more than in the self-interest stories. Children may have either inferred that the character's self-interest caused him to lie or that bias influenced the character in such a way that it caused an honest misperception of the event. Follow-up studies revealed that the causal link between bias and misperception was relatively difficult but nevertheless possible for children to grasp. Thus 6th graders but not 4th graders, 2nd graders or kindergartners attributed incorrect statements to bias brought about through self-interest.

Theories about how mental state information interacts with skepticism are still a matter for debate. For example, the Mills and Keil study discussed above could be taken as support for the view that mental state information is foundational for evaluating claims. The studies presented here however suggest that any such theory must be constructed so that social information can be used flexibly across contexts. For example, studies 1, 4, 6, and 7 present scenarios in which experimenter expectations are ignored in the evaluation of the conclusions of the researchers. However, in studies 2, 3, and 5, participants' understanding of the researchers' mental states did influence the degree to which they adopted the conclusion. This suggests that across different situations, the importance of social information may be weighted differently. Understanding the precise nature of such processes and the precise ways in which informational cues interact remains an important topic for future research.

Practical implications

Meta-analyses in both psychology and the medical field suggest that experimenter bias and false results are a distressingly common occurrence (Ioannidis, 2005; Yank, Rosenthal & Rubin, 1978; Pfeiffer & Hoffmann, 2009). Moreover, the scientific findings reported in the press tend not to be the most reliable (Gonon et al., 2012). By neglecting a potentially important source of bias, people might accept premature conclusions. This is especially likely if reports in the press make scientific results more appealing than they are—simplifying the design of the experiments, ignoring potential flaws and failing to report financial conflicts of interest. One potential remedy to this problem would simply be to improve scientific standards. Given that the amount of attention that this topic has received recently, such an outcome is likely over the long-term. Nevertheless science will never be entirely bias-free. A better understanding of the possibility and likelihood of experimenter bias and financial bias (e.g. grants from motivated parties), coupled with more accurate scientific reports, would allow the public to be more discriminating in its acceptance of scientific results. On such an approach, the public should be made aware that bias is one factor amongst many that should be considered when evaluating scientists' conclusions. While this might lower the public's trust in specific results, such a finer grained discrimination might also help the public maintain trust on the scientific process as a whole.

One might worry, however, that by alerting the public to the possibility of bias in science, the only outcome would be to further damage public trust in science. For instance, the recent accusations of bias on both sides of the climate change debate might have had this effect, irrespective of the truth of the charges.

It could be overly optimistic to think that the public might become educated about the potential biases of individual scientists or single studies, while retaining trust in the scientific process as a whole. Which of these two options—refusing to mention the possibility of scientific bias so as not to tarnish science's credibility, or stressing the possibility of individual failure so that people retain faith in the whole process—would prove superior is an open empirical, practical and moral problem.

In addition to informing the public's perception of science, the findings reported here could also have implications for scientific practice. As mentioned earlier, false results due to experimenter bias are distressingly common. Most are not created due to conscious fraud but are instead due to an unconscious influence of one's theories on the details of how one carries out a particular experiment (Rosenthal & Lawson, 1964). Perhaps the very general inability to recognize this type of influence accounts for its widespread presence. This could be a topic for future study by, for example, examining whether scientists themselves are prone to the same bias neglect as the general public.

Author Note

We thank Alex Shaw, Angie Johnston, Joshua Knobe, Matthew Fisher, and Frank Keil for helpful comments. This study was supported by the Center for the Study of Mind in Nature (Oslo).

References

Baron-Cohen, S., Leslie, A.M., & Frith, U. (1985). Does the autistic child have a

'theory of mind'? Cognition, 21, 37–46.

Cain, D. M., Loewenstein, G., & Moore, D. A. (2005). The dirt on coming clean: Perverse effects of disclosing conflicts of interest. *The Journal of Legal Studies*, *34*(1), 1–25.

Collins, H. M. (1983). The sociology of scientific knowledge: studies of contemporary science. *Annual Review of Sociology*, *9*, 265–285.

Eagly, A. H., Wood, W., & Chaiken, S. (1978). Causal inferences about communicators and their effect on opinion change. *Journal of Personality and Social Psychology*, *36*, 424-435.

Fleck, L. (1981). *Genesis and development of a scientific fact.* Chicago: University of Chicago Press.

Gonon F, Konsman J-P, Cohen D, Boraud T (2012) Why Most Biomedical Findings Echoed by Newspapers Turn Out to be False: The Case of Attention Deficit Hyperactivity Disorder. *PLoS ONE* 7(9): e44275.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk. *Perspectives on Psychological Science*, *6*(1), 3.

DeScioli, P., & Kurzban, R. (2009). The alliance hypothesis for human friendship. *PloS one*, *4*(6).

Heyman, G. D., Fu, G., & Lee, K. (2007). Evaluating claims people make about themselves: The development of skepticism. *Child development*, *78*(2), 367.

Mercier, H., & Strickland, B. (2012). Evaluating arguments from the reaction of the audience. *Thinking & Reasoning*, *18*(3), 365–378.

Mills, C. M., & Keil, F. C. (2005). The Development of Cynicism. *Psychological Science*, *16*(5), 385–390.

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, *5*(5).

Pfeiffer T., & Hoffmann R. (2009). Large-Scale Assessment of the Effect of Popularity on the Reliability of Research. PLoS ONE 4(6): e5996. doi:10.1371/journal.pone.0005996

Sodian, B. (1988). Children's attributions of knowledge to the listener in a referential communication task. Child Development, 59, 378–385.

Wakefield, A.J., Murch S.H., Anthony A., et al. (1998). RETRACTED: Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children. *Lancet, 351*, 637–41.

Wilgenburg, E & Elgar, M.A. (2013). Confirmation bias in studies of nestmate recognition: A cautionary note into for research into the behavior of animals. PLoS ONE 8(1), e53548, doi:10.1371/journal.pone.0053548.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. Cognition, 13, 103–128.

Yank, V. Rennie, D, and Bero, L. (2207). Financial Ties and Concordance Between Results and Conclusions in Meta-analyses: Retrospective Cohort. *BMJ*, *335*, 1202-1205.

Tables and captions

Experiment	Percent female participants
1	62%
2	57%
3	61%
4	51%
5	49%
6	30%
7	32%
8	28%

Table 1: Percent female participants in Experiments 1 to 8

Experiment	Average age	Standard deviation age
1	36,1	12,9
2	35,9	12,2
3	32,3	12,2
4	35,0	12,9
5	32,5	12,3
6	28,1	9,9
7	29,6	9,4
8	26,8	8,4

Table 2: Average and standard deviation of participants' age in Experiments 1 to 8

Experiment	Percent participants with at least some college
1	89%
2	90%
3	87%
4	85%
5	86%
6	92%

7	87%
8	84%

Table 3: Percent participants with at least some college in Experiments 1 to 8