**Reviewer #1 (Remarks to the Author):**

In this manuscript titled "Neural mechanism underlying verification and falsification in human reasoning" the authors examine how humans learn from evidence suggesting a stimulus is an example of a rule ("target") or not an example of a rule ("non-target"). They suggest humans are biased in learning from information suggesting a stimulus is a target (referred to as confirmatory feedback or verification feedback) than from information suggesting a stimulus is a non-target (referred to as non-confirmatory feedback or falsification feedback). Using a computational model they suggest this strategy is not optimal. They relate this bias to striatum and prefrontal activity. The implication is that human scientific inference may be flawed.

The question of whether we learn better from an example of the rule/group or from an example that is not part of the group is interesting. The question of whether we learn better from data that is consistent with our hypothesis vs. data that is inconsistent with our hypothesis (i.e. if we are correct or wrong) is also interesting. However, these are not the same questions and the implication of each is different. While the authors are in fact examining question number one the two are confounded in their study and their writing, and this is the major problem of this manuscript.

In their paradigm a subject may decide a stimulus is a "target" or "non-target". They define "confirmatory" evidence as evidence suggesting a stimulus is a target (pg 4). This is a case if the subject says "target" and is correct OR if the subject says "non-target" and is incorrect. They define "disconfirmatory" evidence as evidence suggesting a stimulus is a non-target (pg 4). This is a case if the subject says "non-target" and is correct OR if the subject says "target" and is incorrect. The problem is that the subjects are biased to saying "target" - they say target more than non-target (Figure 1 B, bottom panel). This means that "confirmatory" evidence is mostly evidence from saying "target" and being correct. "Disconfirmatory" evidence is mostly evidence from saying "target" and being incorrect.

Example: There are 16 trials. Subject says target on 10 trials and non-target on 6 trials. The accuracy rate is at least 50%. Lets say it is 50%. So confirmatory evidence: 5 trials target-correct and 3 trials non-target incorrect. Disconfirmatory evidence: 5 trials target-incorrect and 3 trials non-target correct. As you can see confirmatory evidence is confounded with being correct (i.e. with conformation bias). In the behavioural analysis it seems that the authors disregard this factor all together. It is highly likely that the subjects are not learning better from confirmatory evidence (that something is a target) but rather they learn better from correct feedback. While this is interesting on its own it is not a novel observation (see Westen et al., 2006; Doll et al., 2011). Regardless, the authors must dissociate the two to conclude where the bias is originating.  
For the fMRI study the authors do introduce the factor of error/correct together with the factor of confirmatory/disconfirmatory. However, they do not in fact show differing activation for confirmatory/disconfirmatory while controlling for error/correct. The problem partially lies in the fact that error trials are more likely to be oddballs in the confirmatory condition while correct trials are more likely to be oddballs in the disconfirmatory condition, which could underlie the difference in BOLD activity.

Moreover, while the authors clearly indicate in the results that their interest is in how we learn from targets vs. non-targets (not from correct/error) there are frequent sentences in the abstract, introduction and conclusion that give the impression that they are in fact interested in how humans learn from information that supports priors (for example: "we show that humans participants learn faster from evidence that confirms (rather than disconfirms) a hypothesis" also sees pg 10 second line).

Other concerns:

1. Could the authors clarify the intuition behind disconfirmatory evidence being more valuable in their task? If there are a minimum of 6 possible rules and disconfirmatory evidence roles out two rules (for each stimulus in a pair) than we are left with a minimum of 4 possible rules. However, if we receive confirmatory evidence than we are left with 2 possible rules. This suggests that confirmatory evidence is more valuable, no? I may have missed something along the way, so clarification would be helpful.

2. The manuscript is difficult to understand because of inconsistent use of terms (such as confirmatory, verification..) as well as not an optimal selection of terms. Confirmatory can be misunderstood as confirmation that one is correct, for example. Also the use of "stimulus" to describe a pair is somewhat confusing as the intuition is that a pair consists of two stimulus. A clarifying rewrite that simplifies the methods and results would be extremely helpful.

**Reviewer #2 (Remarks to the Author):**

Summerfield and colleagues examine learning behavior and concurrent fMRI in subjects performing a category learning task. Category membership is determined by a disjunctive feature rule, which means that negative examples are more informative than positive examples of the target category. However, fits of a learning model to behavior show a tendency to learn from positive as well as negative examples, interpreted as a bias toward learning from confirmatory examples; and differences in neural activity across conditions track this effect. I find myself puzzled and unconvinced by a number of aspects of the authors' interpretation. Some of these difficulties may relate to a lack of clarity in the manuscript, or my own difficulties understanding it.

1) It remains unclear to me that the central bias toward confirmatory evidence is a real phenomenon. For behavior, most of the evidence is derived from fits of a computational model whose details are unusual, poorly motivated, and not clearly related to the issues at hand. Most importantly, if I am reading Equations 2.1 and 2.2 correctly, the learning model builds in the artificial constraint that any learning from confirmatory examples must come at the expense of less learning from disconfirmatory examples. (Relatedly, the only way to adjust the stepsize of learning, bounded by the magic leading constant 1/2, is to push the fixed squared-stepsize budget from one outcome type to the other.) Since these two sorts of feedback happen on different trials, I don't see any reason that this needs to be the case.

But the prediction that alpha should be near 1 (all learning should be disconfirmatory) seems clearly wrong, and like a side effect of optimizing under this unfortunate constraint, since allowing any learning from confirmatory examples would have to come at the cost of reducing learning from disconfirmatory examples, and the latter are more informative by design. A more reasonable model would allow a separate stepsize parameter for each sort of evidence; presumably in this case optimal behavior would include both sorts of learning even though one is more helpful than the other. Subjects who behave this way, when fit with the more constrained model, would exhibit alpha < 1. So I don't think that this parameter estimate tells us that subjects exhibit a suboptimal bias.

2) More generally, the behavioral model seems ungrounded and unmotivated, and the test of it against optimal Bayesian inference is insufficient to justify its details. I don't understand the motivation for considering min H(i,j) as a possible decision variable in Equation 1, or what this tradeoff really has to do with the hypotheses about confirmation-based learning. An obvious standard approach would be a linear perceptron / Rescorla-Wagner type model, which responds on the basis of the sum (equivalently, mean) over H. In the two-hypothesis case (familiar cues) I think this actually corresponds to tau=0.5 and close to the fit result. Also, the categorization literature (which is most closely related to this task) has a long tradition of associative learning models. Many issues posed by this task have been studied there, including credit assignment (e.g. the Rescorla-Wagner account of cue competition) and dimensional attention (e.g. Kruschke's ALCOVE and later variations such as his 2001 J. Math Psych paper), but these possible mechanisms are not considered.

3) The other behavioral finding that may bear on the existence of confirmation bias is Figure 2D. I don't know what to make of this. The main finding seems to be that the slope of p(target) by evidence count is larger for confirmatory than disconfirmatory evidence. But this should be trivially true of most any model (even for the authors' optimized model with alpha near 1, I would think) because disconfirmatory evidence should decrease, while confirmatory evidence should (if anything) increase, target responses: the nature and meaning of such evidence guarantees that these lines slope in opposite directions. I don't really know how the relative steepness of these effects should ideally be expected to vary if they are optimized independently, but it seems like this test is just driven by the difference in slope direction anyway.

It is further shown that this slope effect differs for novel vs familiar targets. The finding that some effects differ between these conditions is often taken as further evidence of a confirmation bias in this article. The intuition seems to be that these biases may be related to a sort of working set limitation whereby subjects learn by focusing on one or a small number of hypotheses, giving rise to confirmatory reasoning. This seems like quite a reasonable idea, except that it also seems to have nothing to do with how the authors' actual model works.

3) The neural results, collectively, show patterns of effects that differ as a function of condition. But I had a hard time seeing how these relate to the authors' model, or how they support the idea of a bias toward learning from confirmatory evidence.

For instance: The finding that the model as fit to behavior correlates with vmPFC whereas the reward optimizing one does not is problematic for the reasons mentioned above. Additionally the finding that one contrast is significant while another is not, does not imply that difference in the fits is itself significant.

Next, it's just not really clear what to make of the ROI-based ANOVAs. In 3c, a main effect of confirmatory vs nonconfirmatory feedback types is taken as evidence of a bias toward learning from confirmatory examples. But presumably many features differ between these trial types including EV which is previously argued to be represented there; it's not clear whether this might explain or confound the effect (or whether it would even under a more ideal model like the reward-optimized one). Moreover, in Figure 5a, the finding that a similar pattern of results differs as a function of familiar vs novel cue (which it does not in 3c) is taken as a key signature of the supposed bias. Which is it? Also, leaving aside concerns about the constraint under which alpha is fit behaviorally, it's not clear why the particular contrast compared to it in 5b/c (error minus correct, notwithstanding cue or feedback type) is a relevant neural index of confirmation bias to compare to the behavioral one. There just doesn't seem to be a clear line of reasoning that connects all these findings, the computational model, and the informal discussion of issues like capacity.

Finally, the statistics on all of these ROI analyses appear to be biased by mutiple comparisons in the contrasts used to select them. (See Vul, Kriegeskorte etc.) The article at one point mentions "appropriate post-hoc statistics" but it is unclear what this means or how this could be done. The issue is not just the usual one about post hoc tests following the result of a single ANOVA; the problem is to the extent that ROIs were selected over many multiple comparisons using a contrast (like the ANOVA) that is not orthogonal to the subsequent post hoc test, then the subsequent tests are biased. All of these subsequent results basically need to be subjected to whole-brain multiple comparison correction, or the ROIs need to be selected in a non-circular way.

**Reviewer #3 (Remarks to the Author):**

This is an interesting paper on experience-based rule acquisition using fMRI. There are two relatively novel aspects of the study which make it a possible candidate for publication in Nature Neuroscience.

The first is the use of a task which cleverly allows the authors to elegantly differentiate the contrasting ease of learning from confirmatory and from disconfirmatory evidence. This was done by using disjunctive rules and by employing two levels of advance knowledge of what form the rule could take. The study thus built on many psychological studies of confirmatory bias and provided fMRI evidence of the reality of the phenomenon.

Second, the fMRI results are consistently interpreted in terms of what is plausibly argued to be the currently most appropriate purely behaviourally-based model of the process of experience-based rule acquisition. Thus the conclusions of the fMRI analysis build on cognitive psychology models instead of merely and more typically offering ad hoc speculations on the patterns of activation produced.

There are a number of points that require attention before the paper could be accepted. The first is the most important.

1. From figure 5 (B and C) it would seems that as far as DLPFC and RLPFC are concerned confirmatory bias only applied in the novel cue condition and not in the familiar cue conditions. The correlations shown in the figure are no lower for disconfirmation than for confirmation for the familiar cue conditions. In figure 4 too activation across target and non-target trials seems similar for familiar cues, if anything being greater for non-targets. Thus it would appear that for familiar cues, confirmation bias does not occur. Whether it does or not in the familiar cue condition needs to be specifically addressed. If the overall pattern is that confirmation bias only occurs in the novel condition then this can plausibly be interpreted as resulting from an information-overload strategy in this type of task. This seems far from Wason's very deep claims about confirmation bias in hypothesis generation. This issue too needs to be addressed.

2. As far as I can tell - it does not seem to be stated - activations shown in figures 3-5 are averages over all trials in a block. However it is known that in other rule acquisition studies (eg Crescentini et al, 2011) activation in the dorsolateral PFC differs greatly before and after rule acquisition. The authors could use related analyses to determine where in the learning process their frontal and parietal effects are occurring.

3. In view of recent controversies on the robustness of fMRI analyses it is important to know how the number of 18 subjects was decided upon. Poline and colleagues have given estimates on the necessary number of subjects for obtaining robust results. From memory 18 seems a little low.

4. (Minor point) The origin of confirmation bias studies in psychology was Wason's (1960?) famous 2,4,6 study. It should be referenced.