



BIROn - Birkbeck Institutional Research Online

Hahn, Ulrike and Merdes, C. and von Sydow, M. (2018) How good is your evidence and how would you know? *Topics in Cognitive Science* 10 (4), pp. 660-678. ISSN 1756-8765.

Downloaded from: <https://eprints.bbk.ac.uk/id/eprint/23210/>

Usage Guidelines:

Please refer to usage guidelines at <https://eprints.bbk.ac.uk/policies.html>
contact lib-eprints@bbk.ac.uk.

or alternatively

How Good is Your Evidence and How Would You Know?

Ulrike Hahn

Department of Psychological Sciences, Birkbeck, University of London
London, WC1E 7HX

Christoph Merdes & Momme von Sydow

Munich Center for Mathematical Philosophy, Ludwig-Maximilians-Universitaet, Ludwigstr. 31, Raum 128,
D-80539München, Germany

Corresponding author: Professor Dr. Ulrike Hahn, u.hahn@bbk.ac.uk

Classification: Social Sciences; Psychological and Cognitive Sciences

Abstract

This paper examines the basic question of how we can come to form accurate beliefs about the world when we do not fully know how good or bad our evidence is. Here we show, using simulations with otherwise optimal agents, the cost of misjudging the quality of our evidence, and compare different strategies for correctly estimating that quality, such as outcome, and expectation-based updating. We identify conditions under which misjudgment of evidence quality can nevertheless lead to accurate beliefs, as well as those conditions where no strategy will help. These results indicate both where people will nevertheless succeed and where they will fail when information quality is degraded.

Keywords: evidence, diagnosticity, reliability, belief revision, accuracy, Bayes

The Problem of Evidence Quality

Many things about the world we cannot observe directly. These range from the mundane ('has the dog stolen the paper?', 'are these symptoms caused by flu?') to the complex, overarching theories of science (evolution, quantum mechanics and so on). The 'knowledge' we believe to have about these things rests crucially on inference. Even what we consider to be 'direct observation' of the world (such as the relative depth of objects in our visual field) is the result of constructive, inferential processes that create a 'model' of the world based on the perceptual evidence available [1].

Any such inference faces the fundamental question of the extent to which a given body of evidence licenses the conclusion we wish to draw. For certain contexts, this question has a clear, normative, answer in Bayesian belief updating, as we will outline below; but the accuracy of such inference will be limited by our understanding of the quality of the data: how well do they indicate the truth or falsity of our hypothesis? Crucially, in many real-world situations these data characteristics will not be known exactly or even not be known at all, as in the case of the stranger giving us directions, the witness in court, or the social media post about a politician's behaviour. What

can we do in these circumstances and what are the implications of uncertainty about our evidence? Specifically, anyone with an interest in accuracy should be deeply concerned with the question of a) how we could come to know the quality of our evidence and b) exactly how the accuracy of our beliefs is compromised where our estimates of the quality of the evidence is wrong. Nevertheless, this question has received remarkably little attention. In the following, we address this question by means of simulations with both naïve and optimal Bayesian agents in a setting where the accuracy of their beliefs is well-defined. Such a Bayesian agent will assign probabilities to degrees of belief use Bayes’s rule to update her beliefs in light of new evidence:

$$P(h|e) = \frac{P(e|h)P(h)}{P(e|h)P(h)+P(e|\neg h)P(\neg h)} \quad \text{Eq. 1}$$

$P(h|e)$ represents the posterior probability of hypothesis, h being true in light of the evidence, e . This probability can be calculated from the so-called prior $P(h)$, and the *diagnosticity of the evidence*: specifically, how likely it is that the evidence would have been observed if the initial hypothesis were true, $P(e|h)$, as opposed to if it were false, $P(e|\neg h)$ (\neg being the logical symbol for negation). This ratio of $P(e|h)$ divided by $P(e|\neg h)$, the so-called likelihood ratio (LHR), provides a natural measure of the diagnosticity of the evidence – that is, its informativeness regarding the hypothesis or claim in question:

$$\text{Posterior Odds} = \text{LHR} \times \text{Prior Odds} \quad \text{Eq. 2}$$

where the respective odds are defined as $P(h)/P(\neg h)$, one calculated before, the other after update in on the evidence. In cases where the likelihoods are known, the Bayesian framework is demonstrably optimal in that alternative inference rules will be less efficient in the sense that they will require larger samples, on average, to be as accurate. Specifically, [2] shows that updating by Bayes’ rule maximises the expected accuracy score after sampling. On average, other update rules will require larger samples to be as accurate. This holds on any measure of accuracy that involves a so-called ‘proper scoring rule’¹ of the kind used to measure the accuracy of probabilistic forecasts, for example, in meteorology [3]. Scoring rules assign credit for correct predictions, and penalties for incorrect ones. Overall accuracy is reflected in the total score. Furthermore, this optimality of Bayesian conditionalization with respect to maximizing accuracy holds not just for ‘interest-free inquiry’, but also holds where actions dependent on our beliefs about the world are at stake: using Bayesian conditionalization to update our beliefs on having sampled evidence maximises expected utility [2]. Finally, [4] demonstrate that for a common measure of accuracy (the so-called Brier score; [5], which is effectively the “mean squared error”), a

¹ A scoring rule is a “proper scoring rule” if it rewards honesty on the part of the forecaster: that is, the expected pay-off determined by the scoring rule is maximal where the forecaster reports her true, underlying belief.

Bayesian will minimise inaccuracy of the agent's beliefs across all 'possible worlds' the agent is conceptually able to distinguish and hence, in principle, to entertain.²

One of the main problems with optimal, Bayesian, inference in the real world, is that the diagnosticity of the evidence, that is the likelihood ratio, *might not be known* (see also [6,7]): How reliably does a piece of witness testimony implicate a defendant? How reliable is the evidence for the anthropogenic origin of climate change? How reliably does a volatile stock market indicate an upcoming crash? How reliably does a change in appetite predict pregnancy? This paper examines the implications of such uncertainty about diagnosticity and the broad classes of strategy we have available for estimating diagnosticity.

(In)accuracy and the Cost of Misjudging the Quality of the Evidence

For our simulations we will use Bayesian agents and the precise definition of 'quality of evidence' they afford, measuring their accuracy across a range of circumstances. Specifically, we will assume that, in the context of the simulation, there is a 'ground truth', and we will measure the accuracy of our simulated agents with respect to that ground truth. Moreover, we will look at aggregate behaviour of our agents over many simulated trials. What we will vary is the true quality of the evidence, the perceived quality of the evidence, and the base rate with which the hypothesis in question is true. To illustrate: imagine that our agents are trying to make a medical diagnosis on the basis of a medical test. The possible disease may be more or less common (base rate), and this may provide a reasonable prior (odds). Moreover, the test may be more or less reliable (quality of the evidence), and eventually at some point in the future the 'true answer' can be resolved (because later symptoms become unmistakable). Obviously, where the medical test is foolproof (that is, it always correctly indicates the presence of the disease, and furthermore never incorrectly suggests the disease is present when it is not), an agent's diagnosis would be 100% accurate. However, where the test is less than perfect (as is generally the case), even the optimal Bayesian agent's 'best guess' will not always be correct. It will, however, represent the best anyone could realistically do. Conversely, an agent who makes a random guess will sometimes be correct. What we are examining is the impact of the agents' beliefs about test quality on their judgment, so that our evaluation must include a baseline of 'the best one could do'. Given such a set-up, we can now examine a question that previous research has (to the best of our knowledge) paid little attention to, namely: what exactly happens to the accuracy of an agent's belief when the agent systematically misestimates the quality of the evidence?

Figure 1a tracks this by plotting the impact on posterior degree of belief as a result of misestimate. The two dimensions of the space represent the true and the subjectively assumed likelihood value; to distinguish these we will refer to the objective likelihood as 'likelihood' and the subjective likelihood as 'trust'. This likelihood value represents the probability of receiving a positive piece of evidence, given that the hypothesis is true (that is,

² With the proviso that these possible worlds are finite, a restriction that seems fine for creatures with finite resources and life spans.

$P(e|h)$).³ For any two values of ‘true’, underlying, likelihood value (y -axis of the plot) and trust (assumed subj. likelihood, x -axis), Fig. 1a shows the difference in posterior degree belief obtained through Bayesian updating (Eq. 1) by assuming that trust value instead of the true likelihood. For example, the point [.8, .6] shows the difference between the posterior degree of belief of an optimal agent assuming the true likelihood of .6 and an agent erroneously assuming .8 instead, given that each has received one piece of positive evidence.⁴ The diagonal running from 0,0 to 1,1 in this space represents the points where true and assumed value coincide, with zero difference as a consequence. In other words, the Figure represents the difference in degree of belief between an optimal agent doing ‘the best one can do’ and an agent who is mistaken about the quality of the evidence, trusting it more or less than she should.

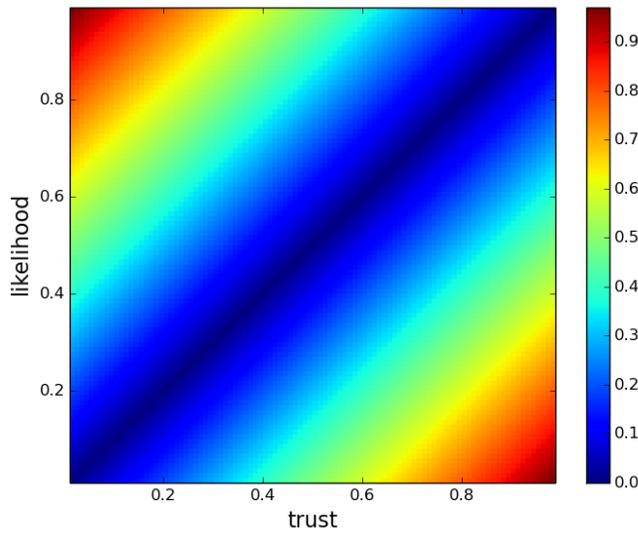


Figure 1A. The figure plots the absolute difference in final (posterior) degree of belief between updating via Bayes’ rule (Eq. 1 above) given one piece of positive evidence and a prior degree of belief of .5, assuming different subjective estimates of the likelihood and the posterior belief obtained using the true likelihood. The x -axis (‘trust’) represents the agent’s subjective likelihood, the y -axis represents the objective, true likelihood.

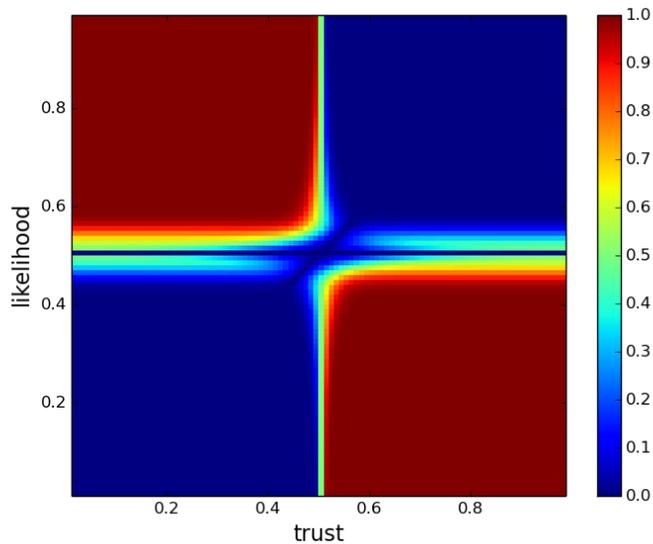
Figure 1A shows that the differences increases with the size of the mis-estimate (that is, greater differences between x - and y -values give rise to greater differences in degree of belief). At the same time, however, mis-

³ For simplicity, we also assume symmetry throughout this paper: agents are assumed to be as good at providing evidence for the hypothesis when it is true, as they are at providing evidence against it when it is false. In other words, $P(e|h) = P(\neg e|\neg h)$, or phrased in signal detection terms, hit rate = correct rejection rate. Note that it does not constrain, in any way, the range of possible likelihoods. Consider the likelihood axis of Figure 1 as it ranges from 0 to 1. For each point on the real number scale in this interval, the likelihood value effectively represents $P(e|h)/(1-P(\neg e|h))$, i.e., a corresponding LHR that runs between 0 at one end (0/1) and infinity as it approaches (1/0) at the other. So in fact the labelling is merely an intuitive way of setting out that LHR scale which expresses the units in something tangible and easy to grasp.

⁴ We assume both start with a neutral prior of .5 This so-called ‘uniform prior’ is widely used as an “uninformative” prior in the sense that it does not express any specific, advance, knowledge about the truth or falsity of the hypothesis in question, thus respecting the so-called ‘principle of indifference’ [8]. This prior also gives rise to the maximum difference in posterior; priors greater or smaller than .5 (and not equal 0 or 1) will shrink the absolute magnitude of the difference without changing the relationships between points in the space.

estimates can be quite considerable and still not be hugely consequential (e.g., mistaking a likelihood of .1 for .25).

Moreover, they become largely irrelevant as more and more data is acquired. Figure 1b shows, for ease of illustration, the same space after 100 pieces of data have been observed, with the composition of that data stream reflecting the expected value determined by the true likelihood. It is a feature of Bayes' rule (Eq. 1) that there is no difference between a series of incremental updates in response to a sequence of independent pieces of data, and between a single update that takes into account all of this evidence at once (see also Eq. 2). Given independent pieces of evidence, the likelihood of the overall sequence is simply the likelihoods of each individual observation multiplied by one another: for example, the likelihood of receiving 3 positive observations of probability .2 is $.2^3 = .008$. As a result, longer and longer sequences of data will move further and further to the origin of the x -axis (i.e., closer to zero). At the same time, differences between the two estimates will shrink due to this multiplication (the difference between $.2^3$ and $.3^3$ is an order of magnitude smaller than the difference between .2 and .3), meaning that differences necessarily erode with more data, as long as the true and estimated likelihood agree on the *qualitative impact*, or *direction*, of the evidence: that is, evidence in favour of a hypothesis has to be viewed as evidence in favour, not as evidence against (in the case of symmetric likelihoods ($P(e|h) = P(-e|-h)$), that means both are on the same side of .5). Flipping what counts as evidence for and against a hypothesis converts a likelihood of p into one of $1-p$ in our plot. In other words, this means that as more and more data are seen, the composite likelihoods (for the entire sequence) are dragged to the respective corners of the plot. Figure 1b shows the same space as Fig. 1a, now after 100 pieces of data have been observed, with the composition of that data stream reflecting the expected value determined by the true likelihood.⁵



⁵ As a binomial process, the expected value is $n \cdot p$.

Figure 1B. The figure shows the same differences between posteriors based on subjective versus objective likelihoods as in Figure 1A, but now after 100 independent pieces of evidence have been seen. The x -axis ('trust) represents the agent's subjective likelihood, the y -axis represents the objective, true likelihood.

This makes clear that what ultimately matters most, and this is the second main feature to emerge from consideration of these spaces, is the perceived *direction* of impact of a piece of evidence: a fever has to be seen as evidence *for*, not against, an underlying flu. Where the relationship between evidence and hypothesis is inadvertently reversed, no amount of data can lead beliefs to converge, as more and more evidence will simply move beliefs further and further in the opposite direction. In this case, we speak of anti-reliable evidence – evidence from, in effect, a systematic liar who we do not know is lying. An agent faced with anti-reliable evidence will only become more and more convinced that the wrong hypothesis is true as additional evidence comes in. Where the qualitative direction of true and mis-estimate match, however, any difference between overconfident, underconfident and correct estimates will necessarily eventually disappear. In other words, given enough pieces of evidence, beliefs will converge, and differences between them will vanish, for different assumed likelihoods (for a proof see Supp Mat B), just as has been shown for different prior beliefs [9].

Strategies for Estimating the Quality of Evidence

Given that there are accuracy costs of varying magnitude to mis-estimating the quality of our evidence, it is worth considering more closely what potential strategies we have for improving those estimates.

The most unproblematic and straightforward strategy is an extensional, frequency-based strategy that monitors the co-occurrence of evidence and eventual *outcomes*: we can come to judge quite accurately the reliability of a pregnancy test, for example, if only we can square instances of test prediction with the eventual outcome that the result predicted. In this way, the diagnosticity of many medical tests is well known (and indeed part of the approval process, e.g., [10]), and statisticians have extensively studied the problem of how best to estimate likelihoods for data.⁶

However, outcome-based strategies will work only where the outcome in question occurs repeatedly and can itself be observed. Many real world beliefs we have do not concern such outcomes. In particular, beliefs we have about singular events (“did Oswald murder Kennedy?”) are not readily amenable to this strategy. Here estimating the diagnosticity of the evidence must itself be based on inference. In a legal trial, we are only concerned with the one case before us, for which we cannot observe directly the ‘true history’, and we will only hear the witness speaking to this one case. What basis might one nevertheless have for assessing the reliability of this witness? One possibility is to try to infer the witnesses’ reliability by drawing on general evidence for which there is some other, indirect, frequency information: for example, we might wish to consider whether the witness exhibits any

⁶ In particular, one may use a maximum-likelihood estimation for the unknown true likelihood p , or a Bayesian inference that incorporates a prior belief about the likelihood to derive a distribution over possible values of the likelihood (see e.g., [11]).

features that have been shown to be indicative of lying in other people [12]. Such an inference itself will provide a less robust and less accurate estimate of diagnosticity than direct observation of large numbers of evidence-outcome pairs, but it is nevertheless grounded, on some level, in observations of past outcomes. Obviously, the space of these broader, more indirect, outcome-based estimates is vast, and the full complement of strategies available in each individual case likely evade complete specification. Nevertheless, Figures 1a and b above give an indication of their *consequences* by indicating the accuracy costs of more or less accurate likelihood estimates. This gives a sense of both limitations and power of all broadly outcome-based approaches.

However, there is a further possible strategy that one might use: assessing the reliability of the witness on the basis of how plausible her statement is. The simple logic of this kind of strategy runs like this: if you say to me something that I think is unlikely to be true, I will nevertheless increase my belief in what it is you are asserting, but I will also decrease my belief in your reliability. On hearing from you that the Earth is flat, this strategy will make me think that this is a tiny bit more likely to be true, but it will also make me think that you are less reliable than I had previously thought.⁷ This strategy not only seems intuitive, but there is also experimental evidence for its use in even very simple contexts of testimony [13]. At the same time, philosophers have considered it to be a rational, normative solution to the problem [6,7,14]. In order to distinguish such a strategy from outcome-based estimates, we refer to this strategy as ‘belief based’ or ‘expectation based’ – because it is simply the mismatch between the evidence expected, given what we presently believe is likely to be true (but do not know to be true!), and the evidence we actually receive that drives the reliability estimate. In the remainder, we examine more closely the utility of this strategy. To this end, we examine the behaviour of a simple Bayesian agent (first proposed by [6]) who treats the match or mismatch between a piece of evidence and his present beliefs about the truth or falsity of the underlying hypothesis as evidence with which to update beliefs about the reliability of the source. In other words, the agent formally implements the strategy intuitively outlined with the ‘flat Earth’ example; full formal specification can be found in Supp. Mat. A.

Testimony and the Reliability of Our Sources

The optimal and naïve Bayesian agents we have considered so far end up with the same posterior degrees of belief whether updating occurs incrementally (after each piece of evidence), or if all of the evidence is taken into account at once (in one single revision of belief). By the same token, the *order* in which evidence is seen makes no difference.

For the belief-based strategy this is no longer the case: because evidence is ‘weighted’ by reliability, its impact depends on the assumed reliability at the point at which it is received. Consequently different individual sequences with the same amounts of confirmatory and dis-confirmatory evidence end up in different places, but

⁷ Note that a strategy that would go even further by ignoring entirely any putative new ‘evidence’ that the Earth is flat would display confirmation bias so extreme that one could never ‘unlearn’ something for which initial evidence had pointed the wrong way.

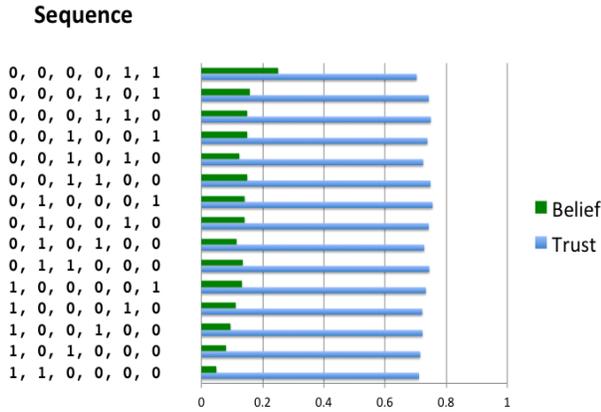
they will also differ from taking the evidence provided by that sequence into account all at once. We thus need to simulate what the average behaviour of such agents will look like.

At the same time, the lack of order independence means that the main condition for belief convergence is no longer met. More and more evidence will not necessarily lead to accurate beliefs (see Supp. Mat. B).

In effect, belief-based updating makes it harder for the evidence provided by a benign world to assert itself. Since testimony is weighted by the extent to which it is congruent with our present belief about the claim in question, the agent exhibits a kind of ‘confirmation bias’, whereby belief congruent evidence becomes amplified, and incongruent evidence down-weighted (e.g., [15, 21]). Sensitivity to the mere *order* in which evidence is received follows from such ‘confirmation bias’ as a consequence, see Figure 2.⁸

⁸ Order effects could only be avoided by the agent either a) specifying in advance a model that foresaw all possible future combinations of evidence the agent might receive and defining an initial joint probability over them, or by b) remembering each piece of evidence and, at each time step, taking *all of the* evidence and recalculating from the initial priors. Neither is realistic in practice, which is why we follow Olsson (2011) in letting our simulated agents update locally on the basis of the new information available at each time step, Past evidence is contained only in the agents’ ‘prior’ at the step.

Panel A



Panel B

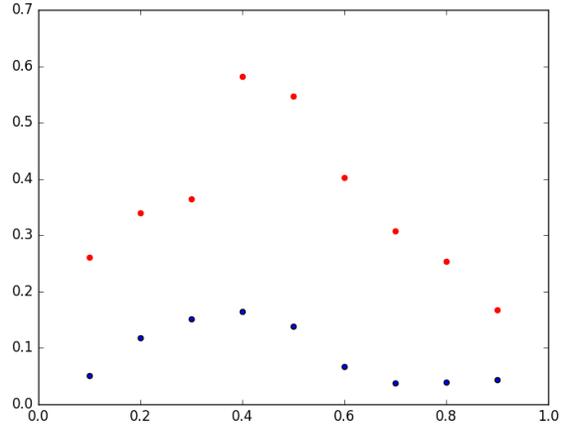


Figure 2. Panel A (left) provides examples of the order dependence for the same 6 pieces of confirming and disconfirming evidence for the trust updating agent. Shown are the posterior degrees of belief in the hypothesis and the expected value of the posterior trust function. The variability in posteriors is sizeable. The simulated results assume a prior degree of belief of .6 and a beta-distribution with $\alpha = 5$, $\beta = 1$ for the trust function, which gives an initial expected value for trust of $p = .83$. Panel B offers a more systematic view of the impact of order effects. The y-axis depicts the maximum difference in posteriors for all possible orderings of six 1's and three 0's, across different prior degrees of belief (x-axis). Red dots represent the difference in the highest and lowest posterior degree of belief observed within the set of sequences, blue dots represent the largest differences in posterior trust. The prior trust distribution of the agent was beta(2,1), which gives an expected value for the initial trust function of $p = .66$.

These features seem problematic: though there is ample evidence of information order effects in our everyday reasoning [see e.g., 16], such order effects are rarely viewed as rational: the very same pieces of information should ideally yield the same conclusions, regardless of the mere coincidence of which were received first and which ones later.⁹ However, merely viewing exactly the same evidence in different orders leads expectation-based updating to different beliefs, both about reliability and the truth or falsity of the claim in question and this should ring alarm bells about such a strategy.

Fuller insight into the behavior of expectation-based updating is shown in Figure 3 below. The figure compares the impact on posterior degree of belief and accuracy (left most column) with that of an agent who simply assumes the source is moderately reliable with a *fixed trust* value of .66 and does not seek to modify this degree of trust in light of incoming evidence. Row 1 shows the (mean) posterior degree of belief (y-axis) of both agents after observing 10 pieces of evidence (averaged across 1000 such agents). As can be seen, that posterior varies both as a function of the true (but unknown) base rate (z-axis) and the true (but again unknown) objective likelihood (x-axis). Consider first the case where the evidence is of the highest quality (likelihood = 1). Here, the evidence is perfectly diagnostic and the fixed-trust agent, shown in Column 2, will converge on the true

⁹ Of course, there may be contexts in which seeming order effects *are* rational (for example, in the context of so-called Jeffrey conditionalisation), but closer analysis reveals that in such circumstances *something further*, beyond the mere order of presentation, has also changed [17].

hypothesis: scanning along the z-axis one sees that where the hypothesis is always false (base rate = 0), the agent will come to a mean posterior degree of belief of 0; and where the hypothesis is always true (base rate = 1) that mean posterior degree of belief is 1. At each base rate level between these two, the mean posterior corresponds to that base rate, reflecting that the agents have correctly identified those cases where the hypothesis is true. At the other end, however, where the source is perfectly anti-reliable (with Likelihood $p = 0$), the inverse pattern holds: false claims are consistently believed to be true! Row 2 shows how these posterior degrees of belief translate into accuracy, measured here with the Brier score --a widely used proper scoring rule (the mean squared error; see also above section “The problem of evidence quality”). In keeping with the posterior degrees of belief just described, the fixed-trust agent is completely accurate (with an error score of zero) at high underlying likelihoods, that is, when the data are consistent and reliable, but maximally inaccurate

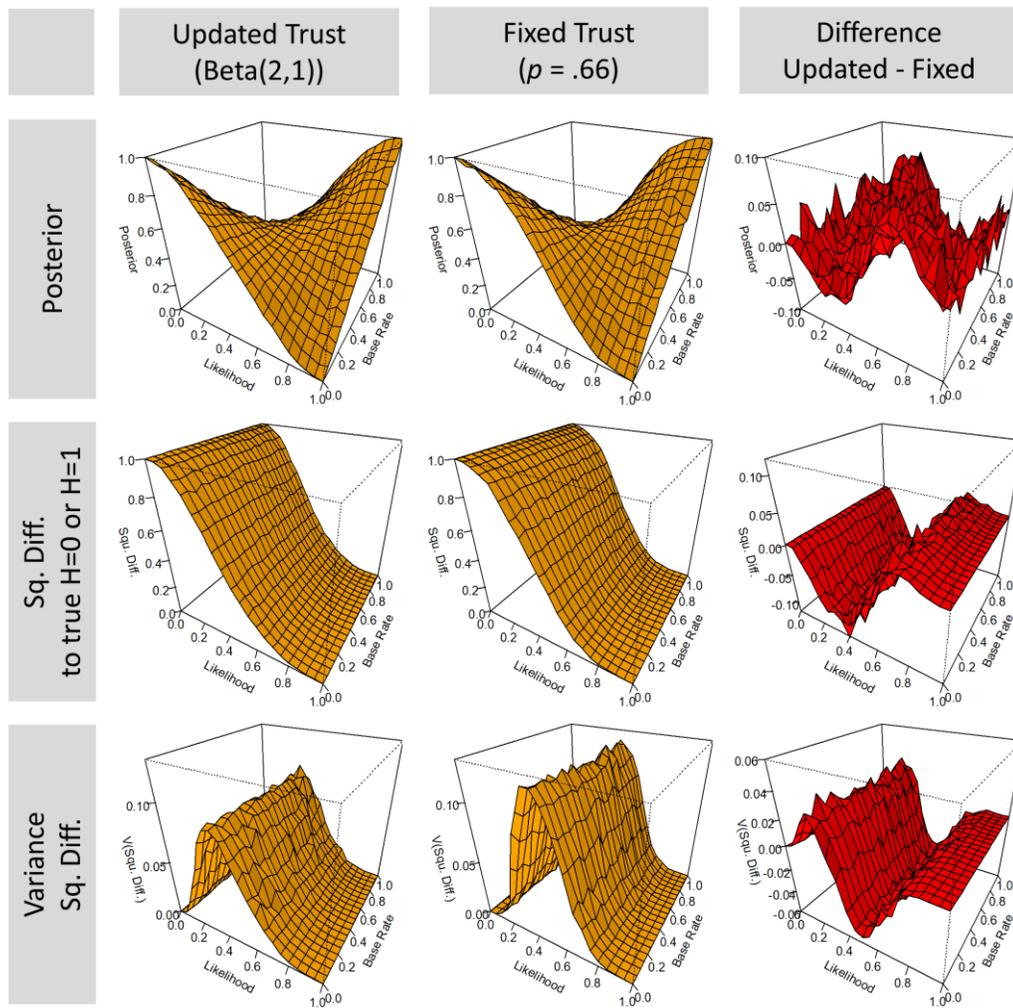


Figure 3. A comparison of fixed-trust agent and a trust-updating agent across the range of possible likelihoods and base rates. Shown are the results of simulating 1000 sweeps per data point of an agent receiving 10 pieces of evidence at the underlying, true likelihood (x-axis), assuming the estimated fixed likelihood of .66 for the fixed-trust agent or, for the updater, an initial beta distribution $\alpha = 2$, $\beta = 1$ (a distribution with an expected value of reliability = .66). In all cases, the initial degree of belief in claim H itself is .5. The y- axis in Row 1 shows posterior degrees of belief in H, the y-axis in row 2 squared error between the posterior and the true status of H, and the y-axis in row 3 the variance thereof.

where the data are consistent but “anti-reliable” (that is at the lowest likelihoods) and intermediate in the middle range where the data stream is highly variable (at likelihood .5, and equal number of instances of “*e*” and “*not-e*” are the most likely outcome). Row 3, which shows the variance in the error scores, shows how the variability of that data stream affects the variance of the error score.

In short, as seen above with respect to Figure 1, where the world dispenses high quality evidence, the fact that the agent mis-estimates the true likelihood has little impact, because with enough evidence the posterior degree of belief will nevertheless converge on the true value. By the same token, the agent is helpless in the face of evidence that consistently ‘lies’. The main question, however, is whether a strategy of belief-based updating can help here. Column 1, Figure 3, shows the corresponding plots for the agent who updates reliability based on the extent to which the data match present beliefs about the hypothesis. The striking finding is how little the performance of this agent differs from that of the fixed trust agent: the posterior plots in row 1 are hard to distinguish with the naked eye, so column 3 provides a difference plot to assist the comparison. Likewise, there is little difference in the error, or for that matter, in the error variance. A moment’s reflection clarifies this counter-intuitive result. One can think of the range of possible (true underlying) likelihoods as encompassing three distinct zones, as shown in Figure 4 below.

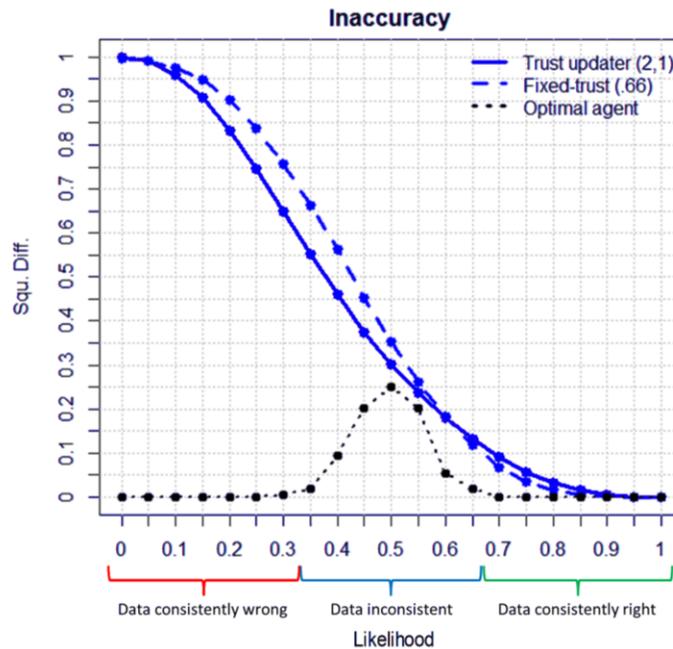


Figure 4. The graph shows the mean (in)accuracy (squared error) averaged across base rates after 10 pieces of evidence. In other words the figure collapses the *z*-axis (base rates) in plots Figure 3B (second row). It shows the trust updater and the fixed-trust agent in a single figure and sets them against the accuracy score of an optimal agent who knows the true likelihood.

In the “green zone” (Fig. 4), the data are consistently ‘right’, in the “red zone” the data are ‘lying’ (anti-reliably providing support for the false hypothesis), but consistently so, and in the middle, “blue zone”, the data stream is highly inconsistent (in the urns example above, at $P = .5$ one would expect an equal number of red and blue balls in the sample drawn). Crucially, at the low likelihoods of the “red zone”, the consistency of the data bars the trust updating agent from detecting the anti-reliability. Because the agent starts from an uninformative prior of .5, the initial trust the agent displays means that the first few bits of data move the belief in the hypothesis below .5 and subsequent data are entirely consistent with this belief, meaning that the agent only becomes ever *more* convinced that the anti-reliable source is, in fact, reliable. As a consequence, faced with a ‘Cartesian Demon’ [18] who systematically “directs his entire effort to misleading” a trust updater is ultimately as lost as the fixed trust agent. The belief-based trust updater can only exploit inconsistency in the data stream to modify its beliefs in the source’s reliability, and in the limit, where that inconsistency vanishes, the performance of the two agents converge. By the same token, the biggest difference between the types of agent is in the middle zone of high data variability (“blue zone”, Figure 4). However, the high variability of the data itself constrains how large a benefit the trust updater can accrue. Basically, precisely because the data are so variable, there is not much that can be learned here – as is apparent from the black line in Figure 4 which represents the error score of a Bayesian agent, who, like fixed-trust and trust-updating agent starts with an uninformative .5 prior, but actually *knows* the underlying likelihoods. Finally, in the “green zone”, where the data are consistent and reliable the updating agent performs less well than the fixed-trust agent, because the fixed-trust agent converges on the true hypothesis, and all trust updating does here is add noise to that convergence process as the update latches on to slight inconsistencies --in a mirror image to the dynamics just described for the other end of the likelihood range, that is, the “red zone”.

In other words, for the most fundamental problem, that of a consistently lying, anti-reliable data stream, belief-based updating is ultimately powerless; and elsewhere gains are limited or performance is actually worse. The underlying stumbling block is that, with an un-informative prior, the trust-updater has only the inconsistency of data stream itself as a vehicle for boot-strapping trust --as is apparent from Figure 5, left panel, which plots the posterior trust landscape that corresponds to the landscape plots of Figure 3.

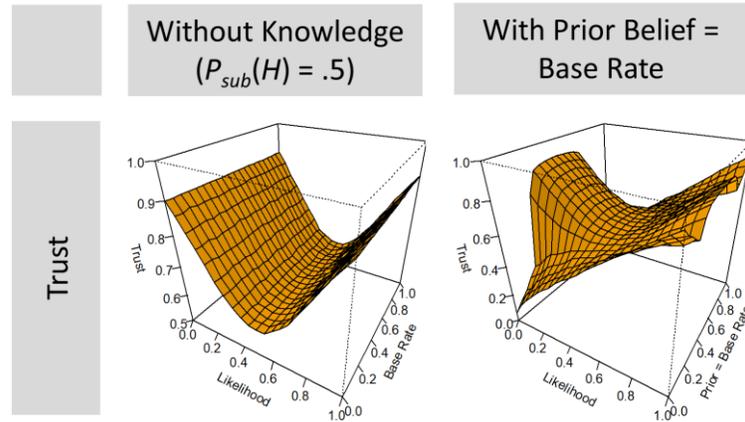


Figure 5. The left hand panel shows the posterior trust (i.e. resulting subjective likelihood) of the belief updating agent in Figure 3. This agent starts with an uninformative prior belief of $P_{sub}(H) = .5$. The right hand panel shows the same agent when that agent knows the true base rate of $H=0$ and $H=1$ and uses this as the prior (cf. next section).

At this point, a reader might be tempted to suggest that the problem stems solely from the fact that, in our simulations, both fixed-trust agent and belief-based updater start out initially trusting. After all, this is the feature that renders both powerless in the face of the Cartesian Demon. The updating agent might consequently fare better if it started with a different level of trust. However, closer consideration of this point reveals that *there is no* ‘better’ starting point: an agent whose initial trust is .5 views evidence as strictly uninformative, so will never learn *anything* from the data. Belief revision will simply not get off the ground. And starting out *dis-trusting*, will simply flip the accuracy plots of Figure 3 and Figure 4 around the .5 likelihood line, leaving the overall level of performance unchanged.

In short, the problems are deep and structural, not merely a result of the chosen parametrisation. The mere fact of data being expected or unexpected in light of our present (uncertain) degree of belief about the hypothesis simply does not convey enough information to solve the reliability problem.

Belief-based Updating Using Prior Knowledge

Things change somewhat once other knowledge on the likely truth or falsity of the hypothesis can be brought to bear. In the simulations thus far, our agents started from an uninformative prior. What happens if, based on other knowledge, the agents start with knowledge of the true underlying base rate?

Figure 6 below shows a corresponding version of Figure 3 for this case. The main difference can be isolated by considering first Row 2: unlike in Figure 3, and unlike the fixed-trust agent, the trust updater now shows a profile that varies across base rate (z -axis). This is because (outcome-based) base-rate knowledge can be harnessed to modify trust as is shown in Figure 5 right plot. As a result, the trust updater can achieve accuracy gains at low or high base rates by becoming dis-trusting of anti-reliable information. This comes at a cost at the other end of the

likelihood range (Zone 1, highly reliable, consistent information), but overall the benefits greatly outweigh those costs. Note also, however, that with this comes increased variance relative to the fixed trust agent (row 3, Figure 6).

Before we move on to draw out the implications of the entire set of simulation results, we discuss the robustness of these simulations.

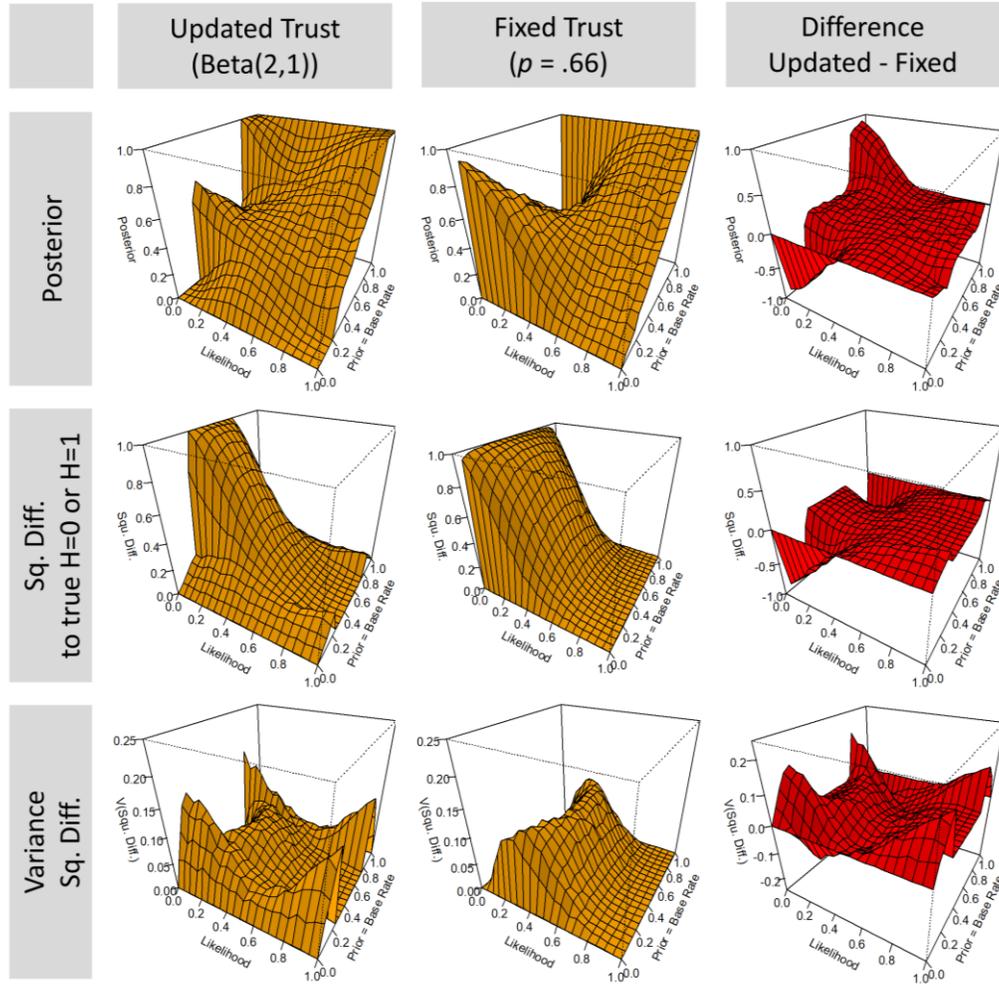


Figure 6. A comparison of fixed-trust agent and a trust-updating agent across the range of possible likelihoods and baserates while assuming prior knowledge by setting the prior equal to base rate. Like in Figure 3 the model was run 1000 times for each data point of the surface and each run involved a 10 step updating process. Row 1 shows posterior degrees of belief, Row 2 squared error between the posterior and the true status of H , and Row 3 the variance thereof.

Robustness and Beyond Bayesian Agents

The obvious question for any reader confronted with a model is the extent to which the results are robust to changes in the modelling assumption. For the interested reader, Supp. Mat. C “Robustness of results” offers discussion of the impact of various simulation assumptions and supplementary simulations. More generally, however, there is ample evidence that human beings are not actually Bayesian reasoners (e.g., [19]) though there is also ample evidence to show that judgments approximate closely Bayesian predictions in many contexts, e.g., [20]; for a review see [21]). One might thus wonder about the relevance of our results to ‘real people’. Here it is important to see that the broad picture that emerges from our analysis is not reliant on agents ‘being Bayesian’. *Any* expectation-based adjustment strategy will show order effects, and any expectation-based strategy will show some form of ‘confirmation bias’ whereby belief congruent evidence is ultimately given greater weight. Likewise, *no* strategy for belief updating will adjust its beliefs in the right direction if it considers reliable evidence to be anti-reliable or vice versa. The key determinants of our results are structural, they do not stem from the specifics of the Bayesian update rule.

At the same time, one might challenge the simplicity of an analysis that is based around a single simple belief p . For one, in the real world, we have a “web of beliefs” which may mutually constrain on another, and so could consequently be expected to help with the problem. From a Bayesian perspective, there are two ways in which other knowledge can be brought to bear: other knowledge may constrain one’s prior degree of belief in the hypothesis at issue or it may constrain one’s estimate of the source’s accuracy. We showed above simulation results addressing **both** of those factors. Concerning the latter, constraints on the estimates of the source, the simulations in the first section of the paper, which show the effect on belief for all possible combinations of objective likelihood and subjective trust, show how much (or how little) better estimates of source accuracy will achieve (Fig. 1a and 1b). The later simulations also allow one to gauge this through comparison of the final beliefs and error squares at varying degrees of difference between the assumed trust value chosen for those simulations ($p=.66$) and the objective likelihood, even though they do not show all possible combinations of objective likelihood and trust. Concerning the first possibility of constraints on prior beliefs, finally, other information one has (including past evidence) that speaks directly to the hypothesis in question is summarised in one’s prior; that is how such information is brought to bear. Our simulations which include appropriate base-rate information reflect that, by providing the agent with the best-possible case of constraining knowledge: namely, knowledge of the true base rate. So the resultant graphs allow one to examine and compare the impact of that ‘best case’ together with the degree of mis-estimate in trust. In that sense, our simulations already triangulate the scope of benefit additional knowledge can have. Consequently, our results suggest that the fact that our beliefs come as broader webs may help, as long as the parts of this web are themselves based on reliable evidence, but this does not over-turn the difficulties highlighted by our simulations. The simulations of Zollman [22], which examine a strategy of ‘subjective calibration’ of trust across a set of beliefs and find it wanting, further underscore this point. Zollman’s simulations differ from the work presented here in that there is no belief revision involved in

his simulations (agents simply add propositions asserted by trusted sources to their stock of beliefs if they do not already have a view on that proposition); nevertheless the results echo ours in finding a fixed-trust strategy not only remarkably competitive, but, in Zollman’s context superior to a strategy of gauging reliability through concordance across a set of agent beliefs.

Implications

Closer inspection of the consequences of not knowing exactly the diagnosticity of our evidence reveals the problem to be more profound than might typically be assumed. At the same time as aspects of the problem are shown to be severe in their implications there are other contexts where mis-estimation really doesn’t alter much: where we are in possession of sufficient data for which we are at least qualitatively right about its impact (reliable vs. anti-reliable) our beliefs will be fine. This explains how it is, for example, that perceptual systems may come to build largely veridical models of the world. Moreover, being right about some things provides the system with a foundation from which the reliability of the system *vis a vis* other things may be assessed on the basis of observed outcomes. This may allow a system to bootstrap itself from finding an initial ‘hook’ into the world to the kind of flexible adjustment to the reliability of sources that the human perceptual system demonstrates in contexts such as multi-sensory perception [23].

At the same time, the fundamental difficulty associated with testimony becomes clear. For large parts of what we believe about the world we are reliant on testimony. This often concerns a *single event* (“Was President Obama part of an illicit ring that met in the basement of a pizzeria?”) for which there can be no outcome based estimates of reliability *vis a vis* that claim. At best, we can seek to establish outcome-based reliability estimates for the reporting source on the basis of other, secondary information (past predictive history, features of the presentation of that claim etc.). However, even that may not be available. In this case, expectation-based updating seems the best we can do. But, as we saw, this mechanism comes at considerable costs of its own while providing only a very limited increase to accuracy. Yet, on average, we need to trust others, because if we do not, we cannot learn *anything* from them.

Understanding the deep tensions involved here seems fundamental particularly at a time when the integrity of our information environment seems under threat. It is undoubtedly important to try to draw attention to secondary features that have at least some diagnosticity in signaling reliability and to increase appreciation of these (e.g., [24]). However, there is a real risk of under-estimating the scale of the challenge and it is our contention that most of us significantly over-estimate human ability to deal with the fact that our information sources are only partially reliable. Expectation-based revision of trust seems so natural as to barely warrant examination. Yet it is a far less powerful mechanism than one tends to assume, and its rationality implications are considerable. If there is a single take-home message from the simulations presented here, it is the extent to which, for better and for worse, we are at the mercy of the integrity of our information sources. Moreover, in the era of fake news, information wars and troll factories, a Cartesian Demon is not just an esoteric philosophical thought experiment. While it is tempting to

attribute what we perceive as ‘outlandish’ beliefs (for example belief in “chemtrails” [25]) to faulty reasoning, the simulations presented highlight the possibility of ‘unlucky’ experiential history: rational agents can come to strongly believe a falsehood, considering its source to be highly reliable; they can come to consider a source to be anti-reliable, thus showing ‘boomerang effects’ whereby they revise beliefs in the opposite direction of what that source asserts [26]; and they can end up with radically different beliefs as a function of the order in which evidence was received. This provides grounds for intellectual humility, but also highlights the need to safeguard the accuracy of our information. The extent to which we live in a ‘benign world’ often matters far more, it seems, than individual adjustments to perceived reliability that we can make. This suggests that action to maximise the underlying information integrity of our world are paramount, and that, at the end of the day, such actions will have to be grounded in outcome-based evaluations. In conclusion, it is our hope that the simulations presented here not only foster a greater appreciation of the fundamental problems of testimony, but that they may inform the search for technical solutions to the pressing problems of information integrity that our societies need.

Acknowledgments

The research reported in this paper was funded through the Humboldt Foundation’s Anneliese Maier Research Award.

References

1. Knill D C, Richards W (Eds.) (1996) Perception as Bayesian inference. Cambridge University Press.
2. Rosenkrantz R D (1992) The justification of induction. *Philos Sci* 59(4): 527–539.
3. Winkler R L, Murphy A H (1968) “Good” probability assessors. *J Appl Meteorol* 7(5): 751–758.
4. Leitgeb H, Pettigrew R (2010b) An objective justification of Bayesianism II: The consequences of minimizing inaccuracy. *Philos Sci* 77(2): 236–272.
5. Brier G W (1950) Verification of forecasts expressed in terms of probability. *Mon Weather Rev*, 78(1), 1–3.
6. Olsson E J (2011) A simulation approach to veritistic social epistemology. *Episteme*, 8(02), 127–143.
7. Olsson E J (2013) A Bayesian simulation model of group deliberation and polarization. In *Bayesian Argumentation* (pp. 113–133). Springer Netherlands.
8. Jaynes E T (1968) Prior probabilities. *IEEE Transa Systems Sci Cyber*, 4(3): 227–241.
9. Le Cam L (1953) On some asymptotic properties of maximum likelihood estimates and related Bayes estimates. *U Calif Pub Stat* 1(11): 277–330.
10. Phillips K A, Van Bebber S, Issa A M (2006) Diagnostics and biomarker development: priming the pipeline. *Nat Rev Drug Discov* 5(6): 463–469.
11. Kruschke J K (2010) *Doing Bayesian data analysis*. Burlington, MA: Academic Press.
12. Porter S, Brinke L (2010) The truth about lies: What works in detecting high-stakes deception? *Legal Criminol Psych* 15(1): 57–75.

13. Collins, P.J., Hahn, U., von Gerber, Y. & Olsson, E.J. (2018) The Bi-directional Relationship Between Source Characteristics and Message Content, *Frontiers in Psychology*, section Cognition.
14. Bovens L, Hartmann S (2003) *Bayesian epistemology*. Oxford University Press.
15. Nickerson R S (1998) Confirmation bias: A ubiquitous phenomenon in many guises. *Rev Gen Psychol* 2(2): 175.
16. Adelman, L., Bresnick, T., Black, P., Marvin, F., and Sak, S. (1996). Research with patriot air defense officers: Examining information order effects. *Human Factors*, 38(2):250–261.
17. Osherson D (2002) Order dependence and Jeffrey conditionalization, unpublished paper available at: <http://www.princeton.edu/~osherson/papers/jeff3.pdf>
18. Descartes R (1641, 2013) *René Descartes: Meditations on first philosophy: With selections from the objections and replies*. Edited and translated by J Cottingham. Cambridge University Press.
19. Phillips L, Edwards W (1966) Conservatism in a simple probability inference task. *J Exp Psychol* 72: 346–354.
20. Peterson C R, Beach L R (1967) Man as an intuitive statistician. *Psychol Bull*, 68, 29–46.
21. Hahn U, Harris A J (2014) What does it mean to be biased: Motivated reasoning and rationality. *Psychol Learn Motiv* 61: 41–102.
22. Zollman K J (2015) Modeling the social consequences of testimonial norms. *Philos Stud* 172(9): 2371–2383.
23. Ernst M O, Bühlhoff H H (2004) Merging the senses into a robust percept. *Trends Cogn Sci* 8(4): 162–169.
24. Wineburg S, McGrew S (2016) Why students can't google their way to the truth. *Educ Week*. Nov. 1.
25. <https://www.theguardian.com/environment/2017/may/22/california-conspiracy-theorist-farmers-chemtrails>
26. Nyhan, B., & Reifler, J. (2010). When corrections fail: The persistence of political misperceptions. *Political Behavior*, 32(2), 303-330.
27. Savage L J (2000[1954]) *The foundations of statistics* (5th edition). Dover Publications.
28. Parsons S (2001) *Qualitative methods for reasoning under uncertainty* (Vol. 13). MIT Press.
29. Schum D A (1994) *The evidential foundations of probabilistic reasoning*. Evanston (Illinois): Northwestern University Press.

Supplemental Materials

Supp. Mat. A: Model Details

1. Model specification

The model represents a single agent updating beliefs on observing evidence, while simultaneously updating its trust into the source of these pieces of evidence as proposed by Olsson [6,7]. The agent starts from an initial belief $P(0)$ which is a point estimate that hypothesis H is true. The trust of the agent is initialized as a beta distribution over the possible levels of reliability.

The simulation model is structured into three distinct processes:

1. Data generation
2. Belief update
3. Trust update

At every time step t , a piece of evidence is generated at random, and the agent updates belief and trust simultaneously. Following Olsson [6,7], the model assumes evidence symmetry: that is,

$$P(e|h) = P(\neg e|\neg h)$$

As a consequence, only a single parameter p , representing the probability of the evidence given the hypothesis, is necessary to characterize the evidence and positive and negative test results are equally diagnostic, i.e., $LHR^+ = LHR^-$ (for discussion of potential limitations stemming from this symmetry assumption see SuppMat “Robustness”).

Belief update is guided by Bayes theorem using the reliability distribution τ resulting in the following equation for a positive report as suggested by Olsson [7]; the expectation τ , as the original derivation shows, takes the position normally served by the likelihood $P(e|h)$:

$$P_{t+1}(h) = P_t(h|e) = \frac{E[\tau_t] * P_t(h)}{E[\tau_t] * P_t(h) + (1 - E[\tau_t]) * (1 - P_t(h))}$$

where the trust is updated for a positive report according to

$$\tau_{t+1}(h) = \tau_t(h|e) = \frac{x * P_t(h) + (1 - x) * (1 - P_t(h))}{E[\tau_t] * P_t(h) + (1 - E[\tau_t]) * (1 - P_t(h))} * \tau_t(x)$$

Note that, as a result of symmetry, the probability of particular sequences of n pieces of evidence is described by a binomial probability $B(n, P(e|h))$.

For, i.i.d. observations it is only the proportions of ‘successes’ that count, ‘order’ is irrelevant

The probability of exactly k successes in n trials is,

$$f(k; n, p) = P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

For $k = 0, 1, 2, \dots, n$, where

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

and the expected value (mean) is np .

Note also that the proportions of expected trials receiving k pieces of positive and $n-k$ pieces of negative evidence are determined by the true underlying parameters ('the world'), not the agent's beliefs about those parameters. The mismatch between these drives the deviation from optimal, which we determine through simulation.

2. Model implementation

The model was implemented in NetLogo 5.3.1, as well as Python 3.5.2 for external validation. Model runs across multiple parameter settings were conducted using the BehaviourSpace delivered with NetLogo 5.3.1.

Code is available here: <https://www.openabm.org/model/5706/version/1/view>

On a given run, evidence is generated using the **random-float** function provided by NetLogo to generate binomially distributed 0s and 1s.

The numerical integration necessary to approximate the expected value of the trust function has been taken from the example of Laputa to generate compatible results. The algorithm is a variant of the well-known trapezoid interpolation rule:

$$\int_0^1 \tau_t(x) \cong \sum_{i=0}^{n-1} \min(\tau_t(x_i), \tau_t(x_{i+1})) * \delta_x + 0.5 * \text{abs}(\tau_t(x_i) - \tau_t(x_{i+1})) * \delta_x$$

where the interpolation distance is for the above simulations chosen to be 1/100, which provides sufficient precision for the parameter configurations simulated. However, for some configurations (e.g. very extreme trust priors) a smaller interpolation distance may be appropriate.

Supp. Mat. B: Convergence of Fixed-Trust Agents

We want to show that a fixed-trust agent, i.e. an agent whose subjective likelihood $P_{sub}(e|h)$, here also called trust, is fixed and can differ from the objective likelihood of the evidence-generating process $P_{obj}(e|h)$, will still approach $P_{sub}(h) = 1$ given the condition that either

$$P_{sub}(e|h) > 0.5 \wedge P_{obj}(e|h) > 0.5 \quad (1)$$

or

$$P_{sub}(e|h) < 0.5 \wedge P_{obj}(e|h) < 0.5 \quad (2)$$

From hereon, we drop the subscript of the subjective probabilities and likelihoods. Elsewhere in the text we simply use p for the likelihood.

Since the latter case is simply established by a symmetrical argument, we will focus on the former. The proof is simply a reproduction of the one offered by Savage ([27] pp. 46-49) for the case of known objective likelihoods. For completeness, we will walk through Savage's proof with some slight changes to the labeling in line with our model and point to the crucial step where the above assumption figures.

Assume a sequence of random variables X_r , representing the outcomes of individual trials, which are independent. From independence and Bayes theorem it follows for the posterior odds that

$$\frac{P'(h)}{P'(\neg h)} = \frac{P(h)}{P(\neg h)} \prod_{r=1}^n \frac{P(x_r|h)}{P(x_r|\neg h)} \quad (3)$$

where n is the number of observations of independently and identically distributed x . For brevity, we write

$$R(x_r) = \frac{P(x_r|h)}{P(x_r|\neg h)} \quad (4)$$

and

$$R(x) = \prod_{r=1}^n R(x_r) \quad (5)$$

What needs to be established is that $R(x)$ becomes arbitrarily large given enough data, formally

$$\lim_{n \rightarrow \infty} p(R(x) \geq p|h) = 1 \quad (6)$$

for any nonnegative real number c . Put otherwise, it needs to be shown that the product of the likelihood ratios almost certainly goes to infinity.

First, we need to handle a few exceptions and corner cases. We assume that $P(h) > 0 \wedge P(\neg h) > 0$ or, informally, non-extremal priors. Second, we assume that there is no x_r such that $P(x_r|h) = 0$, which would result in $p'(h) = 1$, since this would mean that this particular observation implies h logically. Finally, consider the corner case that

$$P(R(x_r) = 1|h) = 1 \quad (7)$$

i. e. that the data is entirely uninformative with respect to the hypothesis, which precludes convergence to certainty in h . While we exclude this case by condition (1), it will be useful to consider as the limiting case for the proof.

With these cases out of the way, we can consider the actual case of interest, where $R(x)$ is nontrivial and nondegenerate. We start by logarithmizing the expression:

$$(8)$$

$$\log R(x) = \sum_{r=1}^n \log R(x_r)$$

and define

$$I = E[\log R(x_r)|h] \tag{9}$$

Then, by the weak law of large numbers, we obtain

$$\lim_{n \rightarrow \infty} P(\log R(x) \geq n(I - \varepsilon)|h) = 1 \tag{10}$$

for any $\varepsilon > 0$ or equivalently

$$\lim_{n \rightarrow \infty} P(R(x) \geq e^{n(I - \varepsilon)}|h) = 1 \tag{11}$$

Thus, our task reduces to showing that $I > 0$ unless we are in the corner case described in Equation 7. Because of Jensen's inequality, which is applicable due to the convexity of $-\log x$

$$I = E[\log R(x_r)|h] \geq -\log E\left[\frac{1}{R(x_r)}|h\right] \tag{12}$$

For the limiting case of uninformative observations, the expectation of $R(x_r)$ equals 1, and therefore

$$-\log E\left[\frac{1}{R(x_r)}|h\right] = 0 \tag{13}$$

Since for informative observations the right hand side of inequality 12 increases, $I > 0$ is established. This is also the exact point where the assumptions on fixed trust enter. The proof uses not the accuracy of the likelihood, but only the fact that the expectation of likelihood ratios given the data is greater than 1. Given the assumption of our model that

$$P_{sub}(e|h) = 1 - P_{sub}(e|\neg h) \tag{14}$$

that condition given h holds exactly when condition (1) holds. Therefore, the fixed-trust agent converges to $P_{sub}(h) = 1$ whenever this condition holds.

Supp. Mat. C: Robustness & Supplementary Simulations

We discuss here the implications of a number of modeling decisions and consider the robustness of the overall results. In several cases, the figures supporting the point made are sizeable as they summarise many additional simulations. We thus present all of the conceptual points together; the supporting figures follow, one by one, in a single figures section (“Figures robustness”) after the text.

Individual assumptions to be examined:

1. Symmetry assumption

Motivation for symmetry assumption, and limitations of the symmetry assumption: Symmetry means that the evidential impact of a positive and a negative piece of evidence are equal and, in overall effect, cancel one another out. Relaxing symmetry, simply relaxes this constraint. In the absence of specific information about sensitivity and specificity, as is the case in our problem by definition, assuming symmetry seems appropriate, reflecting a different type of ‘indifference’.

It also enables straightforward Bayesian updating, which would not be possible if sensitivity, $P(e|h)$, and specificity, $P(\neg e|\neg h)$, were free to vary, as there would then be two parameters to adjust in response to each data point, leaving the expectation-based revision problem under-determined. One of the two values would thus have to be fixed. At the same time, this would change only the relative impact of positive versus negative evidence, but crucially not the range of LHR’s that are possible for an agent to entertain. As set out in Footnote 3 of the main text and by Eq. 2, this means also that all posterior degrees of belief are possible as the LHR is the only quantity that is relevant for this. Relaxing symmetry would consequently not bring more generality in this regard. It would also not alter any of the structural challenges of the trust update problem as set out in Figure 4 main text.

2. Parameters of the Beta distribution

The beta distribution, $beta(\alpha, \beta)$, used to update trust is characterized by two parameters which determine the shape of the distribution. Across possible parameters, one may use values of α and β that correspond to the same expected trust values ($EV = \alpha/(\alpha+\beta)$) but differ in the variance, ($var = \alpha * \beta/(\alpha+\beta)^2 * (\alpha+\beta+1)$). As our comparisons between belief updaters and fixed-trust agent seek to keep all other things equal, the expected value of the beta distribution specifying the update agent’s initial trust distribution and the value of trust p assumed by the fixed-trust agent should be matched. This leaves free only the variance of the beta distribution: given a fixed ratio, lower absolute values for α and β mean greater variance. Greater variance in turn means greater responsiveness in trust update: less evidence is required to change perceived trust. $Beta(200, 100)$ approximates a fixed trust agent within the bounds of evidence considered in our simulations. The choice of $beta(2, 1)$ for our simulations thus selects an update agent who differs strongly from the fixed trust agent, while still being moderately committed.

Figures S2 below explore the generality of our findings across different epistemic trust values, with $EV = \alpha/(\alpha+\beta)$ of .55, .66, and .83, while keeping $\alpha + \beta$ identical. The results for a trust of .66 (either as a trust prior for the updater, or a fixed value for a fixed-trust agent) correspond to previous results reported in Figure 3 and Figure 4. They have already been discussed. The graphs show that we additionally obtain highly similar results for trust = .83. However, note some differences when the expected value of trust value approaches .5. For trust = .55 there is clearly *higher* error given an objective likelihood of 0, and slightly *lower* error for an objective likelihood of 1. However, there remains a strong similarity between the results for the updater and the fixed-trust agent even for trust = .55.

Figure S2 fixes the ratio to $\alpha/(\alpha+\beta)$ to .66 in all cases, but uses different sums (2,1; 10,5; and indefinite for the fixed trust agent). Figure S2 illustrates that increasing the sum of $\alpha+\beta$ leads the update agent to approximate the results of the fixed trust agents as just discussed.

Overall the Figures indicate that there is no parameter setting that provides a ‘magic bullet’ or fundamentally alters the results concerning the relationship between updating and fixed trust agent reported here.

3. Beta versus other distributions

We chose the beta distribution as it is widely used in the context of Bayesian belief revision (valued both for its flexibility and computational simplicity as a conjugate prior to the binomial distribution). Even though the Beta distribution is the most natural distribution for the ‘ground truth’ determined by a Bernoulli process in our simulations, other distributions for modeling the trust distribution in the update agent would be possible. Thus it cannot, strictly, be ruled out that a distribution exists for which the performance of the updating agent would be significantly improved. We welcome further research on this issue, though we nevertheless think it unlikely that a different distribution would lead to radically different general conclusions. This is because the difficulties faced by both updater and fixed- trust agent are ultimately general structural challenges as evidenced by Figure 4, main text: for consistently reliable information, updating will do little to help, for highly inconsistent, low quality data, uncertainty will remain high for any trust updating strategy sensitive to the data, and consistent ‘lies’ can reliably be identified by recourse to something *outside* of the data stream itself, not by the coherence between data and hypothesis in question.

4. Run length

The accuracy of an agent revising its beliefs about a hypothesis in response to evidence is necessarily affected by the characteristics of that evidence. This also implies that the amount of evidence seen matters as multiple, independent pieces of evidence taken together give rise to a single, composite piece of evidence (see Eq. 2, and Figure 1a and b main text). It is consequently important to probe the extent to which the conclusions reported in the main text are sensitive to the particular run lengths (i.e. number of pieces of evidence) we chose for the simulation. The performance of a Bayesian agent for whom the pieces of evidence are independent is bracketed by two extremes: no evidence (prior = posterior) and so much evidence that it is maximally diagnostic (i.e., LHR approaches either 0 or infinity, except for entirely non-diagnostic evidence in which case prior = posterior as well). Supplemental Figures S3a-c replicate the key results of main simulations (Figure 3, main text) for varying amounts of evidence ranging from 1 to 100 pieces of evidence (as Fig. 1b main text illustrates, the fixed trust agent will be close to convergence at this point). As can be seen from these comparisons, none of the basic conclusions about the fact that expectation-based updating brings little gain (and important costs with respect to order effects) are dependent on the specific simulation run-length we chose for the main manuscript.

5. Other types of repeat interaction

Consideration of different run length in plots S3a-c also allows us to address the robustness of another aspect of the model: what if, instead of multiple repeated interactions with the same person, an agent has brief interactions with multiple people? Would the updating agent fare better under these circumstances? In other words, do greater benefits to trust updating emerge if instead of, say, receiving 100 pieces of information from the same individual, the updater receives 2 pieces of information from 50 agents? Because trust is updated only *after* belief updating, 2 pieces of evidence is the minimal unit for which the trust updating mechanism is efficacious. Considering plots for 2 pieces of evidence in Figure S4 allows one to gauge how interactions with 50 (or 100, or 1000, or 1,000,000 etc.) independent individuals will fare. To see this, note that because the individuals are independent, whatever the perceived diagnosticity of their respective step-2 testimony is taken to be, that step-2 testimony is i.i.d. across individuals. This means the order in which those 50 (or 100, or 1,000 etc.) individuals are interacted with is irrelevant (even though for each of these 50 the data order of the two pieces of evidence they provide *does* matter). As a result, updating sequentially after the testimony of each of these 50 leads to the same outcome as combining all 50 independent testimonies into one single, aggregate likelihood. Consequently, having 50 (or 100, or 1,000 etc.) such interactions corresponds simply to moving to a different objective likelihood on the x -axis of Figure S4. For example, a set of 50 individuals dispensing 2 pieces of positive evidence is equivalent to a single individual dispensing 50 pieces of positive evidence at the same level of step-2 trust. This means the corresponding error score can be read off simply by moving to the appropriate value for this composite evidence. Because independent pieces of evidence combine multiplicatively, this means moving toward either end of the scale (except where the objective likelihood is .5 exactly). In other words, the limits of $p = 0$ and $p = 1$ on the x -

axis representing the likelihoods represent the bounds on the possible cumulative error of an agent engaging with such an array of testimony. As can be seen from Figure S4 at the anti-reliability end ($p = 0$) the updating agent does marginally better (at most by about .025), whereas at the reliable end (toward 1) the updating agent does marginally worse. In short, expectation based updating also fails to provide an adequate solution if one considers aggregates of arbitrarily many independent agents who are interacted with only briefly.

6. Heuristics beyond Bayes

Our simulations make use of optimal Bayesian agents and naïve Bayesian agents with limited knowledge. Given the normative basis for Bayesian models, this sheds important light on the problem people face in the real world. However, real people are arguably not fully Bayesian [18]. One may consequently wonder about strategies and heuristics people do use, and whether there are heuristics other than the ones explored here (and expectation-based update, fixed trust *are* heuristics, albeit ones implemented with an optimal inferential procedure) that might fare better. This, we think, is an interesting question for future research. Once again, however, it is important to realize that the fundamental limitations the updating agent faces stem from structural aspects of the problem (see Figure 4 main text, and item 3 above). We consequently think it unlikely that exploration of other heuristics will do anything other than emphasise the main conclusion that outcome-based reliability estimates are both more powerful and less disruptive than expectation-based processes. At the same time, it is important to remember that order effects appear to be inevitable consequence of any dynamic expectation-based reliability estimating procedure that fails to retain *all* the evidence seen and recalculate posterior degrees of belief over this entire set. We think it extremely unlikely that any heuristic human beings actually use for this problem will do so. We consequently think that the main conclusions hold beyond the specific models examined here.

7. Beyond a binomial world

To evaluate the performance of our models we use a simple world of binomial probabilities. Of course, in the real-world, many data generating processes do not have this structure. It is thus important to consider how the models examined might fare in other contexts. The models examined are in no way intended to be limited to the binomial context, rather they have been proposed and used in past literature (e.g., [6]; see also [14]) as models which act implicitly as if the generating process involves independent pieces of evidence, generated from a stationary environment, even if the setting is one where these requirements are not met, simply because nothing more is known about the data-generating process. Their implicit assumptions thus derive from ignorance or agnosticism about the underlying generative process, and the idea behind them is that they are reasonable models that can be broadly applied where the generative process is unknown. They are *not* proposed as models ideally suited to a binomial process (see Footnote 7, main text). So could there be other evaluation contexts in which the expectation-based updater would dramatically outperform the fixed trust agent? One such context definitely exists, namely a world in which the true likelihoods shift and magically track the subjective likelihoods of the updating agent at every step in time. Here the model's inductive bias matches perfectly the generating process, and by definition no other model could do better. In the real world, such a context is unlikely to exist. What does exist in the real world, are data generating processes that have dependencies between pieces of data (for example witness reports that involve information sharing) and/or that are non-stationary, reflecting a process that changes over time. There is no definitive way of knowing whether or not belief-based updating would suddenly do tremendously well in such a context other than by carrying out specific simulations. However, we can see nothing in the nature of the agents we examined here that would inherently lead one to expect very different patterns of relative performance in other environments.

8. Beyond probabilities

Finally, one might take an even more radical stance toward the generality of our results by querying not just the particular Bayesian heuristics we examine here, but by querying the fundamental framework. Though Bayesian inference is demonstrably optimal in many well-defined settings (see [2,4]), it is by no means consensual that one should think of uncertain beliefs about everyday facts (in particular, everyday facts about singular events such as

“Oswald shot Kennedy”) as representable by probabilities (for extensive discussion see e.g., [28, 29]). The challenge for any such view lies in even defining the basic problem: how is the accuracy of beliefs to be measured and performance to be scored? In the absence of worked out answers to this more basic question, a challenge to the present results cannot even be articulated.

Figures robustness

S1. Other values for the beta-distribution: Changed Means

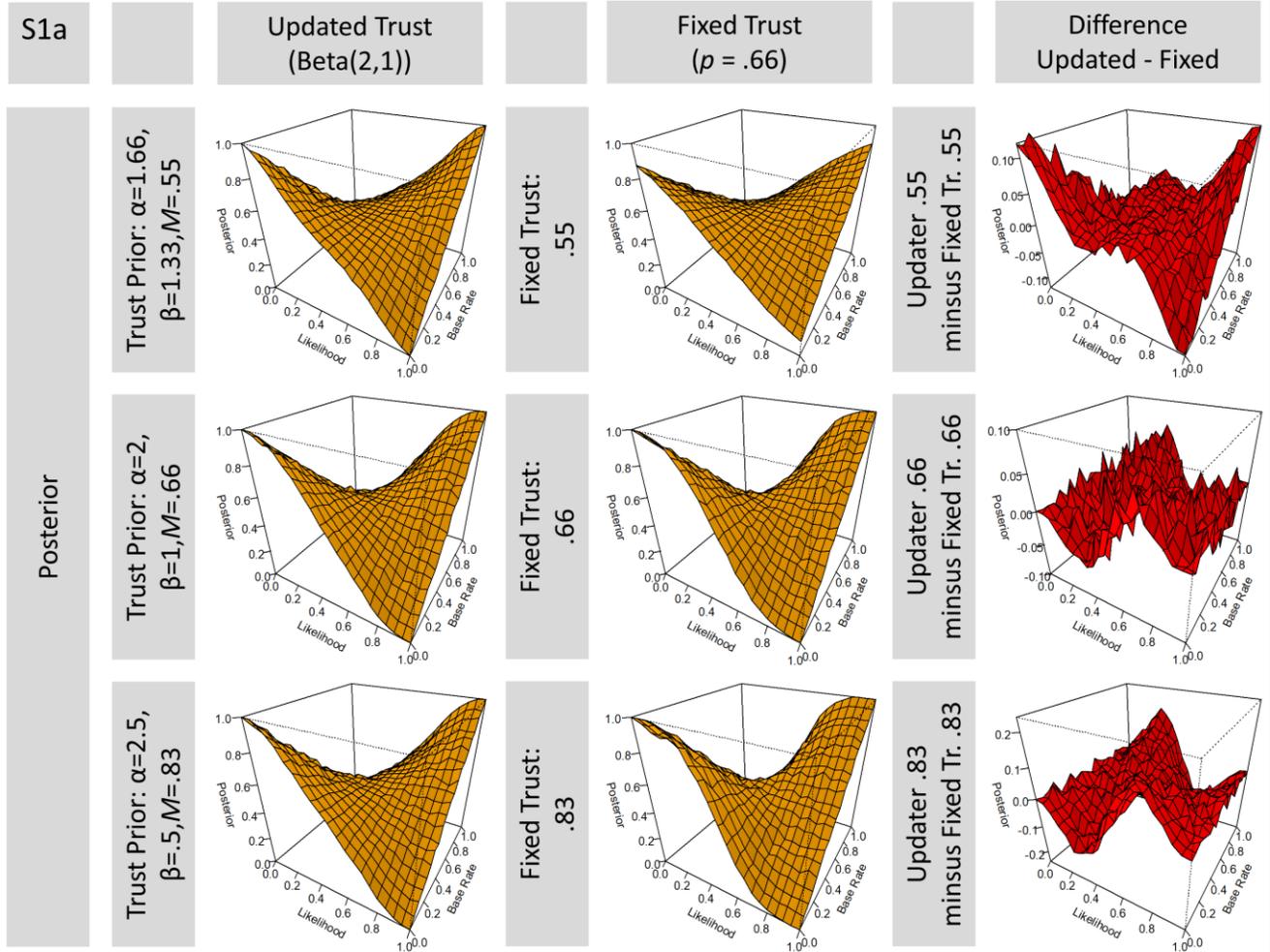


Figure S1a: As outlined in Section 3 of Supp. Mat. C. (parameters of the Beta distribution) the Figure compares different subjective-likelihood (trust) priors and corresponding fixed-trust values (.55, .66, .83), matched across update and fixed trust agent. The graphs show posteriors for the trust updater (Column 1), the fixed-trust agent (Column 2) or their difference (Column 3). For the updater the expected value of the Beta distribution was varied ($EV = \alpha/(\alpha+\beta)$) while keeping $\alpha+\beta$ identical. Each graph provides results for 21 different objective likelihoods and 21 different base rates resulting in represented 441 values for each dependent variable on the y-axis. Since each result is based on 1000 runs of the model, each graph is based on 441000 runs.

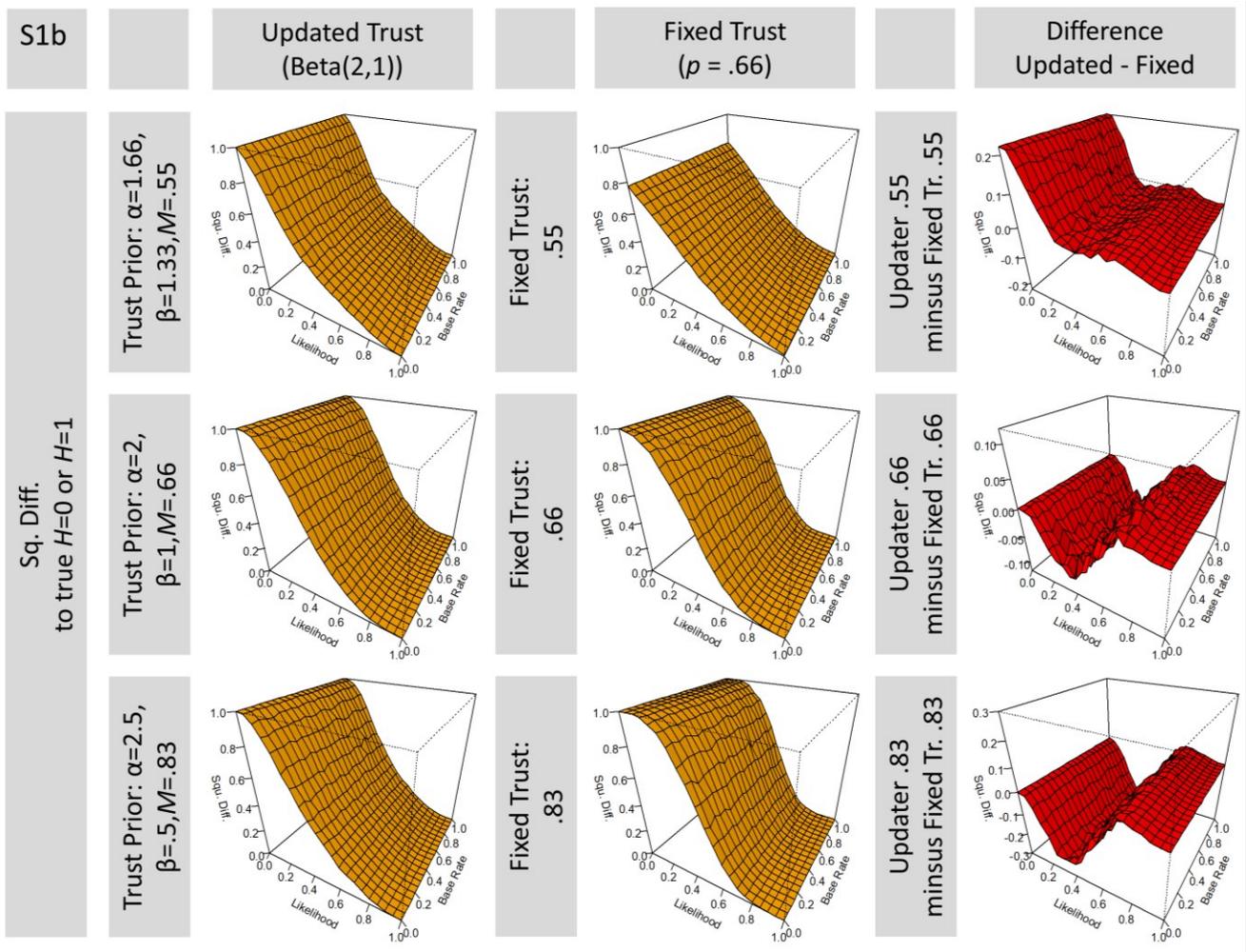


Figure S1b: the Figure compares different subjective-likelihood (trust) priors and corresponding fixed-trust values (.55, .66, .83), matched across update and fixed trust agent. The graphs show accuracy for the trust updater (Column 1), the fixed-trust agent (Column 2) and their difference (Column 3). For the updater, the expected value of the Beta distribution was varied ($EV = \alpha/(\alpha+\beta)$) while keeping $\alpha+\beta$ identical. Each graph provides results for 21 different objective likelihoods and 21 different base rates resulting in represented 441 values for each dependent variable on the y-axis. Since each result is based on 1000 runs of the model, each graph is based on 441000 runs.

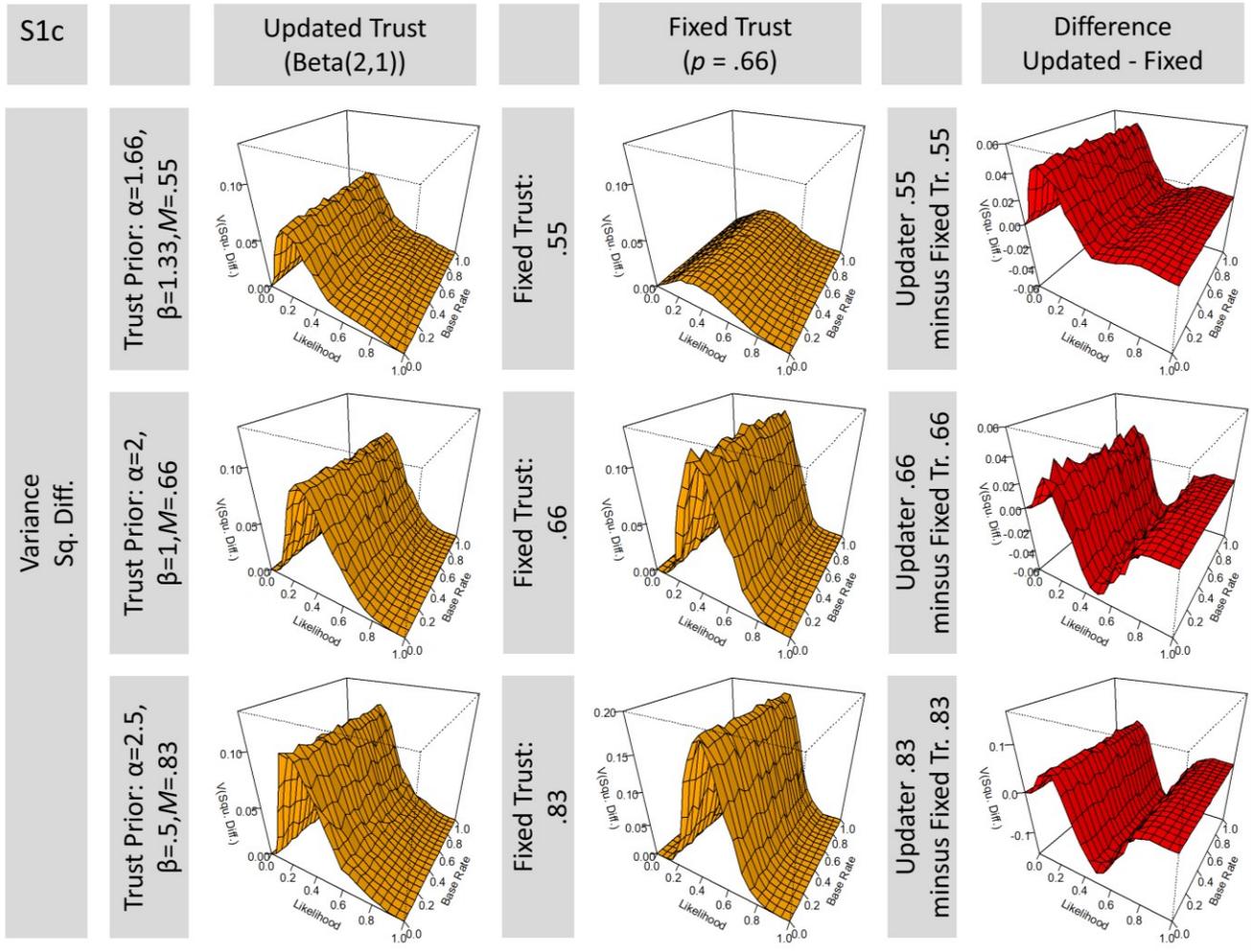


Figure S1c: the Figure compares different subjective-likelihood (trust) priors and corresponding fixed-trust values (.55, .66, .83), matched across update and fixed trust agent. The graphs show accuracy variance for the trust updater (Column 1), the fixed-trust agent (Column 2) or their difference (Column 3). For the updater the expected value of the Beta distribution was varied ($EV = \alpha/(\alpha+\beta)$) while keeping $\alpha+\beta$ identical. Each graph provides results for 21 different objective likelihoods and 21 different base rates resulting in represented 441 values for each dependent variable on the y-axis. Since each result is based on 1000 runs of the model, each graph is based on 441000 runs.

S2. Other values for the beta-distribution: Alternative variance of Beta-distribution

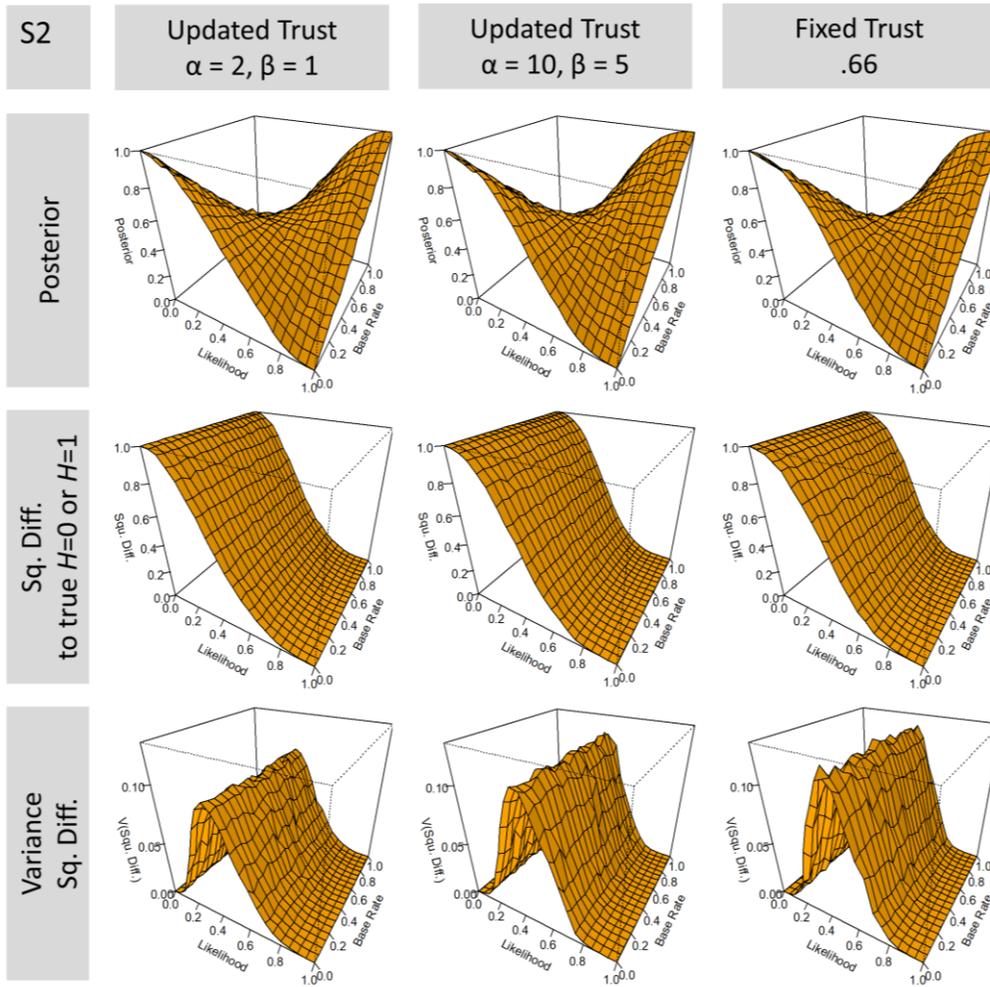
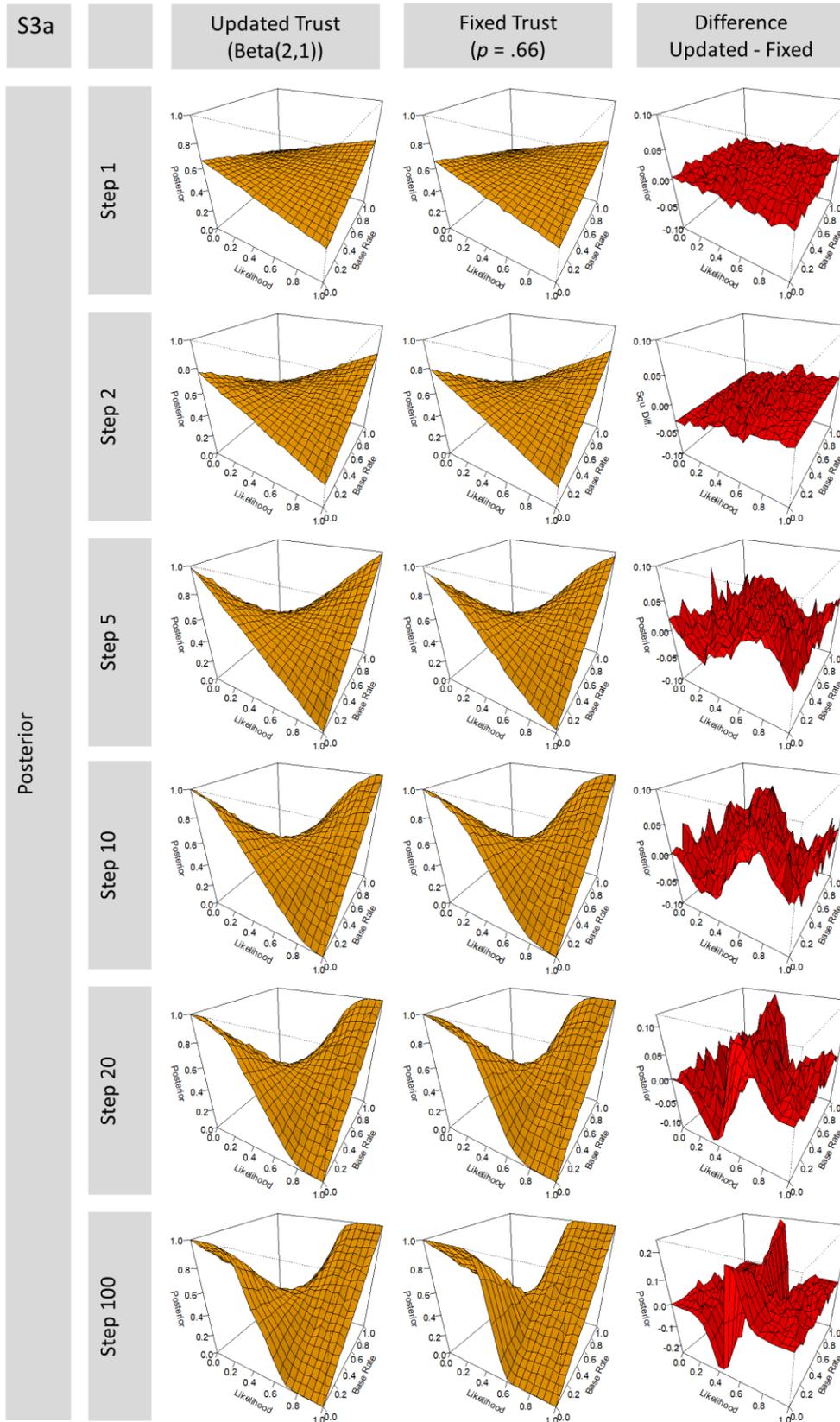


Figure S2a: As outlined in Section 3 of Supp. Mat. C. on the parameters of the Beta distribution, this figure shows increased $\alpha+\beta$ values (with α, β being 2, 1 in Column 1; 10, 5 in Column 2; and fixed in Column 3) with each figure showing results for different objective likelihoods and base rates, and either for the posteriors (Row 1), squared difference (Row 2) or variance of these differences (Row 3). Each graph is again based on 441000 model runs.

S3. Other run lengths



S3b

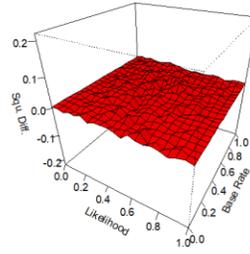
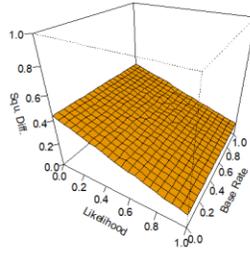
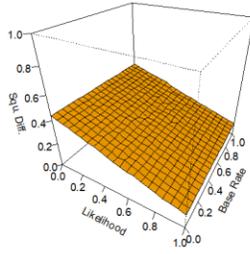
Updated Trust
(Beta(2,1))

Fixed Trust
($p = .66$)

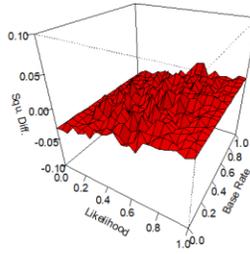
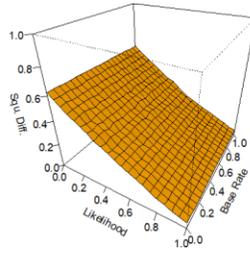
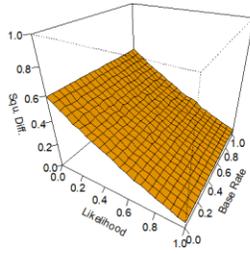
Difference
Updated - Fixed

Sq. Diff.
to true $H=0$ or $H=1$

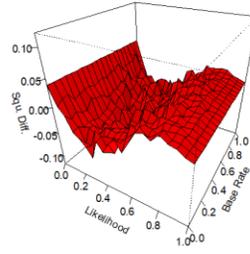
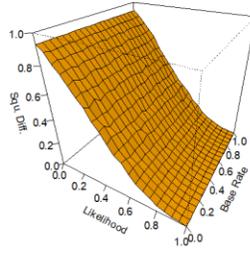
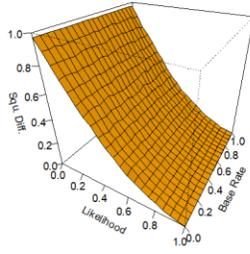
Step 1



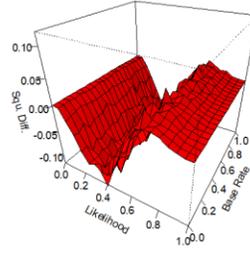
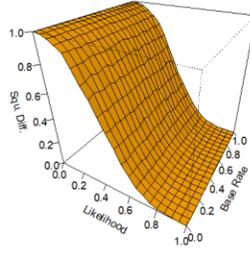
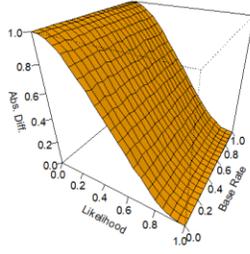
Step 2



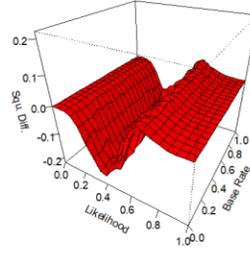
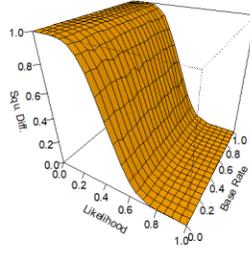
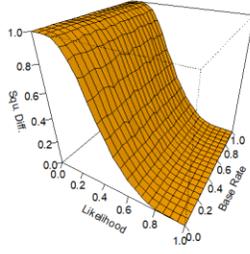
Step 5



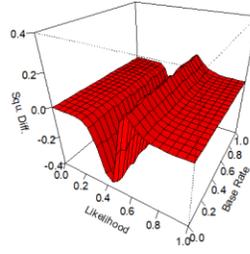
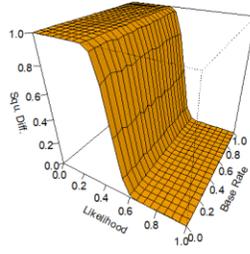
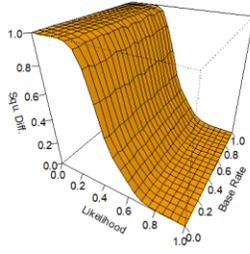
Step 10



Step 20



Step 100



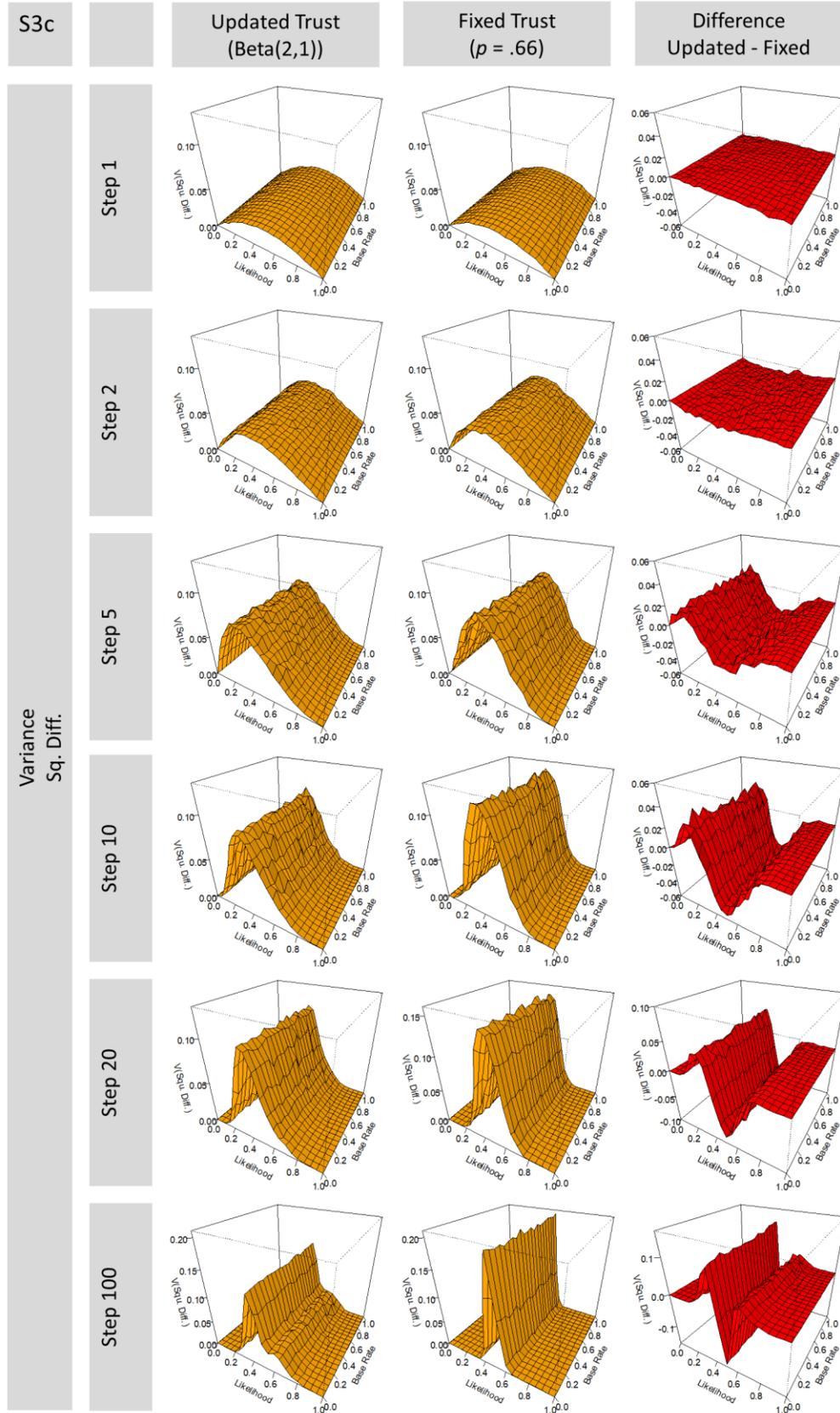


Figure S3 a,b, c: Results analogous to Figure 3 for successively increased numbers of updating steps. The graphs provide a comparison of fixed-trust agent (first column) and a trust-updating agent (second column) with a difference graph (third column with red surface plots). Each graph shows results for a range of possible likelihoods and baserates. For each graph the model was run 1000 times for each data point of a surface (resulting in 441000 runs per surface) and each run involved a 10-step updating process. S3a shows posterior degrees of belief for 1 step, 2, 5, 10, 20 and 100 Steps. S3b provides analogous graphs for squared error between the posterior and the true status of H , and S3c for variance thereof.

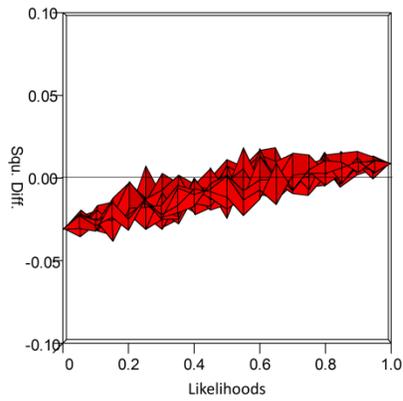
S4. Updating

Figure S4 – Difference graph of updater minus fixed-trust agent for two pieces of evidence. The plot shows the difference in (mean) squared differences at different (obj.) likelihoods, with trust values of .66. Positive values indicate greater accuracy by the fixed-trust agent, negative values greater accuracy for the update-agent. The Figure is marginalized over all base-rates in Figure S3b. Figure S4 is used in Supp. Mat C5.