Why presentiment has not been demonstrated

D. Samuel Schwarzkopf

Experimental Psychology, University College London, United Kingdom

Recently, I published a short Opinion article entitled "We should have seen this coming" (http://journal.frontiersin.org/Journal/10.3389/fnhum.2014.00332) in *Frontiers in Human Neuroscience*. This was a response to two articles by Mossbridge et al. [1,2] in which they discuss so-called presentiment effects, that is, physiological activity (galvanic skin responses, BOLD activity, heart rates, pupil dilation, EEG, etc.) measured before the onset of a stimulus that seems to predict whether or not the future stimulus will be emotional. Taken at face value, such correlations seem to indicate a reversal of causality because future events seem to influence the past. In that light, these effects are similar to another study by Daryl Bem published in a major psychology journal in which he claimed to have found evidence that participants can predict future events [3]. In fact, the Mossbridge article even speculates that the presentiment effects may be a physiological correlate of conscious precognition.

My commentary was of course limited in scope. In it I raised several specific questions about these presentiment studies. More importantly, however, I argued a much more general point about the philosophy of science and how the Mossbridge articles (and much other psi research) in my opinion fail to apply the scientific method. The scope of the article did not really allow me to go into much detail on either aspect although I believe my argument was reasonably clear. However, lack of detail may result in further confusion and it is certainly useful to support my questions with data to illustrate that they are realistic concerns. So here is a list of clarifications of my points as well as some additional ones I couldn't fit into the published article:

1. "Stop relying on statistics at the expense of objective reasoning"

At the end of my article I suggested that statistical inference, while obviously necessary for most scientific research, is insufficient for doing what science is ultimately attempting to do, that is, to derive the best possible models about the world. Statistical inference only assigns probabilities but it doesn't ever *prove* anything (and, as I also pointed out, neither does science itself). This comment could perhaps be misinterpreted as me saying we shouldn't use statistics at all. However, this is clearly not what I said.

Frequentist statistics as used by Mossbridge et al. only tell us about the probability that the effect they observed could have occurred at random. It does not tell us *anything at all* about how probable it is that presentiment (or precognition or any hypothesis) is real. Bayesian model comparison is slightly better because it is a way to assign probabilities to particular hypotheses. In this case, it could tell us how much more probable it is that these effects are actually there relative to the null hypothesis that these effects are random flukes. While this is arguably a better inference, you would still be wrong if you believed that this *proves* presentiment. Any statistics like this can only tell us that *the observations show a reliable difference*. They do not tell us whether these effects actually reverse causality or what other underlying reason there may be for them. Put bluntly, *statistics can never tell us what a result means.*

2. Presentiment invalidates commonly used baseline correction methods

Assuming presentiment effects are really a retrocausal phenomenon, the baseline correction applied in many analyses is clearly incorrect. *If we're lucky,* they only make us underestimate the post-stimulus response (I am working under the hypothesis made by Mossbridge and colleagues that pre- and post-stimulus responses show the same direction). However, *if we're unlucky, an incorrect baseline can completely obliterate a post-stimulus effect.* Many physiological phenomena (for example hemodynamic responses in fMRI, light scattering measured with optical

imaging, as well as many electrophysiological phenomena) are characterized by very subtle differences that would be completely obscured if there is a similar difference in the baseline levels. Conversely, in some situations the opposite may be true as well: if there is a difference in pre-stimulus responses that is unaccounted for by our baseline correction, this may introduce an artifactual post-stimulus difference that isn't really in the data. If there is a positive pre-stimulus difference but in truth there is no post-stimulus difference, by not taking the pre-stimulus difference into account we will end up with an artifactual negative post-stimulus difference.

If presentiment really exists, who knows how many post-stimulus effects are artifacts? In order to know this we will have to reanalyze the entire body of evidence in neuroscience. As opposed to what some may say, I am not afraid of this. However, I think it is correct to demand that before we do this we should know if this effort is indeed really necessary.

3. The quality of the primary data

A significant result in a meta-analysis is completely meaningless if the data it analyses is of poor quality. In my commentary, I mentioned one of the studies included by Mossbridge and colleagues, a functional MRI experiment claiming to show presentiment effects [4]. As I described, this study commits several major errors that received a lot of attention by the fMRI community in recent years: multiple comparison correction and circular inference (double dipping). Inadequate adjustment for false positives can have dramatic consequences in neuroimaging because it typically performs statistical tests separately for tens of thousands of voxels. One ironic study demonstrated this problem [5] by showing that lax multiple comparison correction reveals significant responses to a cognitive task in a dead fish (I feel a bit uncomfortable mentioning this study in the context of a discussion about psi effects…). Circular inference on the other hand is when the same data are used for selecting and making statistical inferences on an effect [6,7]. This is a form of over-fitting a model to random noise which inflates the estimated size of the underlying effects.

In addition, one point I didn't raise in my commentary, the experimental design also does not really account for the sluggishness of hemodynamic responses. Both the pre- and post-stimulus periods had fixed durations of 8.4 seconds. This is not exactly a rapid event-related design but it is too short for what the authors were trying to achieve. In rapid fMRI designs it is critical that trial orders are counterbalanced in such a way to maximize the efficiency of the statistical comparisons. Moreover, it is imperative to jitter the timing of each trial as this helps estimate the signals for different experimental conditions reliably. This study took neither of these steps. Probably a better (albeit inefficient) design would have been to use really long trials of around 60 seconds. A typical hemodynamic response takes up to 30 seconds to decay back to baseline levels. With a really long trial it would have been possible to estimate true pre-stimulus activity. One problem with such a design is that slow drifts in the signal could confound the signal time course but this is a smaller problem and could have been addressed analytically.

In all honesty, these are however only two concrete criticisms I can make for this Bierman study. It has many other weaknesses. The figures in that article are actual screenshots of the analysis package, showing time series for particular voxels apparently showing the effect of interest. They are of such poor image quality that it is impossible to discern any useful detail. While some spatially normalized coordinates for the effects are reported, it is entirely unclear what these brain regions are, why they are supposed to be relevant, or how they were chosen. On top of that the writing is very poor and in parts completely incoherent. Every single undergraduate essay I have read as part of my teaching duties has been clearer than this study. Quite frankly, when I read this study I thought it was a joke. I would be surprised if any respectable journal in the field would have even sent this manuscript out for peer review.

I singled out this particular study because it uses fMRI, a technique with which I am more familiar than with skin conductance measures or some of the other modalities analyzed in presentiment studies. It may seem a bit unfair to focus only on this study. It is true that removing this from the analysis would hardly influence the overall results. However, that is not the point. The issue is not that this is a bad fMRI study. It does not take an fMRI expert to spot that this study is of such poor

quality that it seems mind-boggling it was included. It is essentially the science publication equivalent of a grainy photograph of a UFO or Bigfoot sighting. It is impossible to verify the information in that article or to derive any meaningful evidence from it.

My point in mentioning this study was that *it casts doubt on the entire database used in the meta-analysis.* I admit that some of the other studies I read are undeniably of better quality (including some by the same author of this fMRI study). However, a lot of these studies are difficult to obtain. Many are only published in conference proceedings or are otherwise not fully published. The question I am asking is this: if Mossbridge and colleagues included one study of such obvious low quality, *how many other similarly poor studies are there in the database?*

4. Imbalanced ratio of control and target stimuli

Another issue I pointed out in my commentary is the imbalance between the number of control and target stimuli used in typical presentiment experiments. Frequently it is close to 2:1 although some studies use ratings of the emotional valence of individual stimuli which may alter the actual ratio. Mossbridge and colleagues briefly discuss this imbalance but they don't go into much detail on whether this could be a potential confound. It is a critical problem for several reasons.

First of all, the fact that the two stimulus classes do not appear with equal probability completely undermines the main premise of the studies espoused by the titles of both the Mossbridge articles. *The events are not unpredictable!* After only a few trials an attentive participant will have noticed that targets appear less often than control stimuli. If participants guess that the next trial will contain a control stimulus, they will on average be correct two thirds of the time (Figure 4.1). This seems pretty predictable to me.

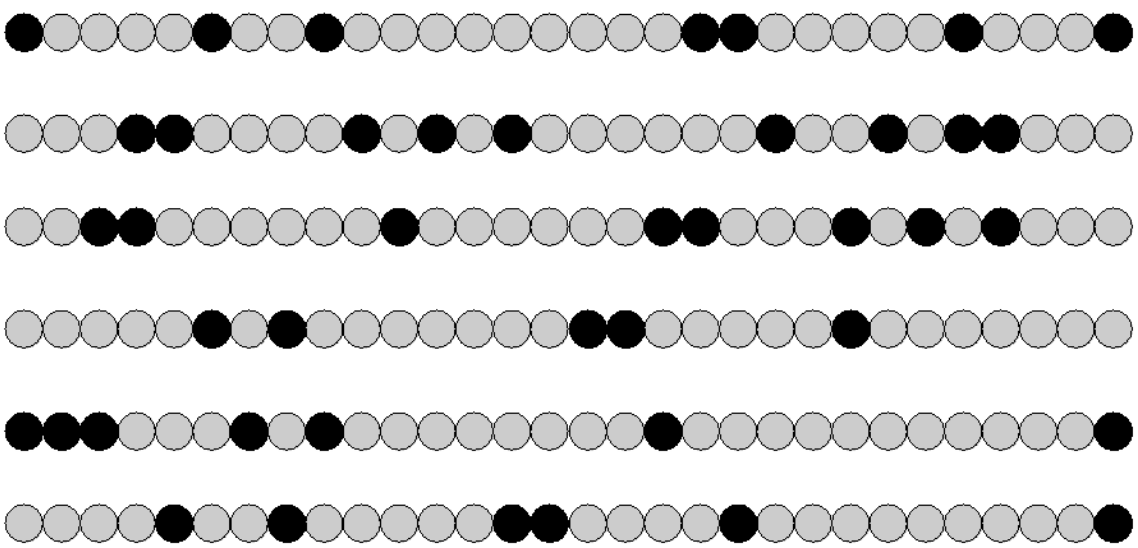

Figure 4.1. Six random sequences of target (black) and control (grey) trials randomized with a 2:1 probability of being a control. The statistical regularity in these sequences should become obvious to any attentive participant after only a few trials.

This predictability must influence the expectations participants will have and thus also be reflected by physiological responses. As I discussed in my article, many presentiment studies perform analyses trying to detect potential expectation bias. However, in most cases I have seen (certainly some of the better presentiment studies in my own judgment) this is limited to a control analysis in which the pre-stimulus effect is correlated with the length of time between target

events. However, because the number of long sequences of control trials is exponentially smaller than that of short sequences the reliability with which the signal for long sequences can be estimated is much worse. Because the pre-stimulus activity is subject to variability this in turn means that the statistical power to detect a linear relationship between the pre-stimulus activity and the number of control trials must be very poor (Figure 4.2).

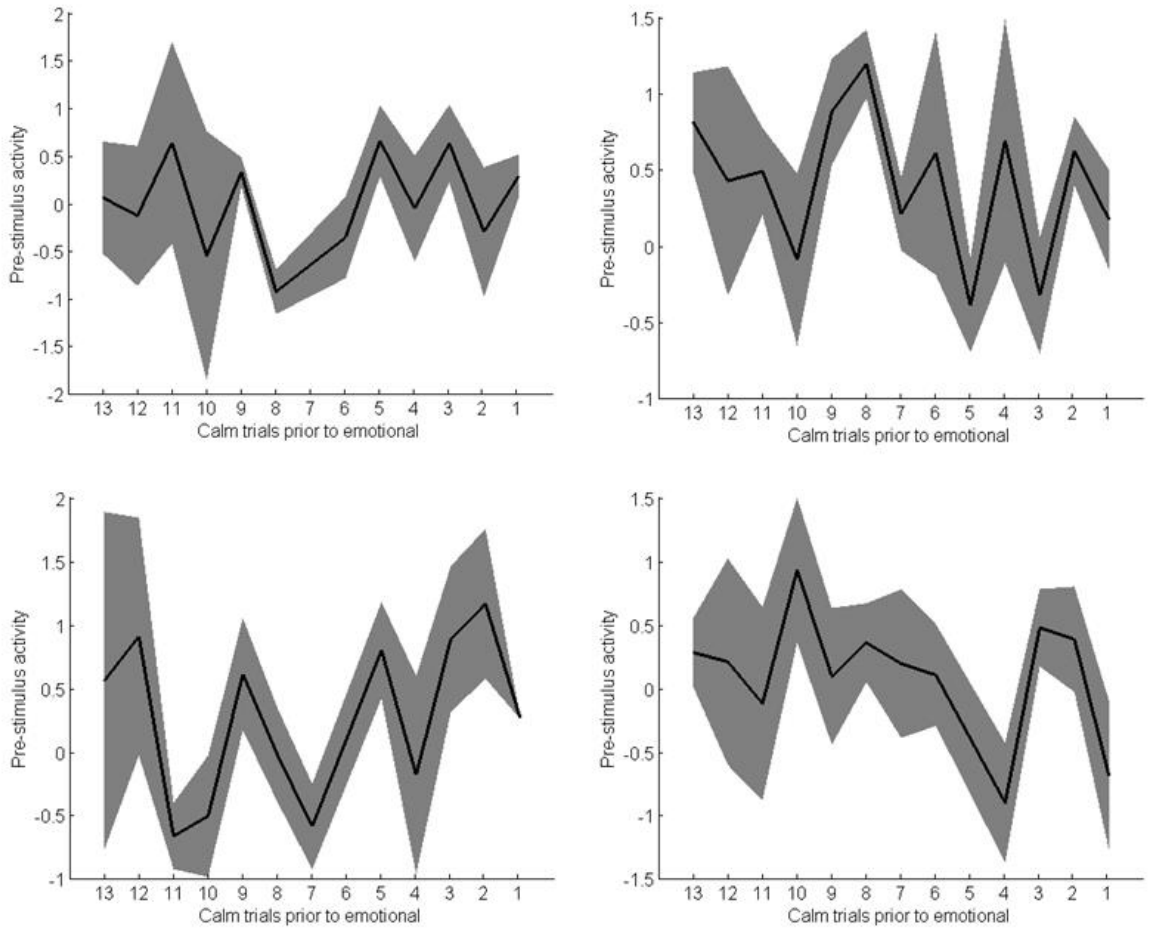

Figure 4.2. Commonly used control analyses may be insufficient for detecting a linear expectation bias effect. A weak pre-stimulus signal which increases linearly with the number of calm trials prior to an emotional trial was simulated using the parameters (same number of participants and trials, ratio of trial types) of a published presentiment study. Four typical examples of this analysis show no obvious linear relationship between the number of calm trials prior to an emotional trial and the pre-stimulus response. The plots are in fact very similar to the one in the original study [8]. To estimate the statistical power of this approach I performed this simulation 10,000 times. A linear correlation was significant on only 3.74% of simulations.

More importantly, who ever said that expectation effects scale linearly with the strength of expectation? To me this seems to be only one of many possible expectation effects. In fact, I wouldn't be surprised if pre-stimulus activity to expectation would remain largely constant from trial to trial. In that case this analysis approach can't possibly work.

Participants presumably also learn about the statistical regularities of the sequence. With a 2:1 ratio it is four times more likely that a control trial is followed by a control trial than that a target is followed by a target. Longer strings of targets are even less probable. This will quickly shape the pattern of expectation for a participant. This means that while after a stretch of several control trials there is likely to be an expectation effect. After a target trial a participant will be more likely to expect a control. Of course this is a gambler's fallacy as the true probability that the next trial will be a target is in truth always the same. But the human brain isn't very good at detecting the

actual probability (especially not after a handful of trials). It is however very good at detecting patterns – even when there isn't one.

Expectation effects may also be further distorted by adaptation effects (such as when signals become progressively weaker over a series of control trials). Alternatively, adaptation in itself may play a role in producing these effects. Since long stretches of control trials are more likely than those of targets, it may only take a few pairs of target trials to produce a subtle difference in the pre-stimulus period.

5. Potential artifacts introduced by "clamping"

I also raised the question of "clamping," an analysis technique typically used by presentiment studies. Clamping means that the physiological measurement at a particular sample (usually the first) is subtracted from all the samples in an epoch. Of course, this means that after clamping the response at this time point will therefore always be zero and everything else is relative to this single point. The problems inherent with this approach should be immediately obvious: it makes *all* the data dependent on this *single data point*. A principle fundamental to all data analysis is that the more data one collects the more reliable are the summary statistics one can estimate from it. Baseline correction is no exception of course. Ideally one should have baseline periods comprising the *same amount of data* as the signals one tries to estimate. This rule can be relaxed at times but the more balanced the signal estimate is to the baseline the better. In fact, in many situations the best thing to do is to calculate a *common baseline* for the entire data set (possibly after detrending and removing of slow drifts and fast temporal noise). In functional MRI experiments what is typically done is to correct an entire time series relative to the global mean across all conditions (including the baseline trials). In electrophysiology it seems common to epoch each trial separately. However, it would be useful to include null events in which we don't expect any response and estimate the baseline from that.

At the very least, however, a whole period should be used for estimating the baseline. In normal experiments, where researchers do not expect to find retrocausality, one would probably use all the data up to stimulus onset to compare the post-stimulus response. Obviously, this baseline period is invalid for testing the presentiment hypothesis. However, it is clearly insufficient to use only a single data point as baseline because even if this point is a true baseline (i.e. it is invariant to the different conditions), by virtue of being a single point it introduces greater variance into your estimation than a longer baseline period. Naturally, if one uses a large number of trials and/or participants, this variability should eventually cancel out. Nonetheless the problem remains that larger variability means lower reliability. With many of the presentiment experiments I would therefore be doubtful that the signals were estimated reliably.

As I discussed in the commentary, this becomes an even greater problem when the single data point used for clamping is *not a true baseline*. If the measure at that particular data point is still affected by the response in the preceding trial, you will inadvertently introduce artifacts with this normalization procedure. Note that this problem is not specific to clamping. Any normalization to baseline using an incorrect baseline estimate will introduce artifacts. So even with standard methods the preceding trial could influence the results. However, clamping vastly exacerbates this problem. Imagine for example that the signal after a strong response (to an arousing target stimulus) is larger and decaying at a different rate than after a calm control stimulus. This is not unrealistic (in fact it can be seen in many signal time courses shown as illustrations of the presentiment effects in the literature, e.g. [8]). In turn this means that there is a temporal dependency between the point used for clamping and all the subsequent data points in the pre-stimulus period.

This temporal dependency can translate into a larger pre-stimulus signal after an arousing target than a calm control. This is where the imbalanced ratio of controls and target trials comes in again. Since most trials are preceded by a control trial, we can expect greater adaptation to control trials than the odd-ball targets. This makes significant differences in the pre-stimulus period much more likely when they precede the rare targets than the frequent controls. Note that standard baseline correction (i.e. without clamping) removes this artifactual difference in most

cases (Figure 5.1A) because all the data have been corrected by the mean and variance in the pre-stimulus period already. However, with clamping it is quite common for there to be significant differences in the pre-stimulus period (Figure 5.1B). I simulated this 1000 times for stylized time courses like those shown in Figure 5.1. In this particular case the percentage of simulations with significant pre-stimulus differences as hypothesized by these presentiment studies is close to 30%. The exact rate of this depends on the magnitude of the post-stimulus signal difference between trial types. In my simulation this is arguably smaller than in many real presentiment studies, so this simulation is likely to provide a conservative estimate of this problem.

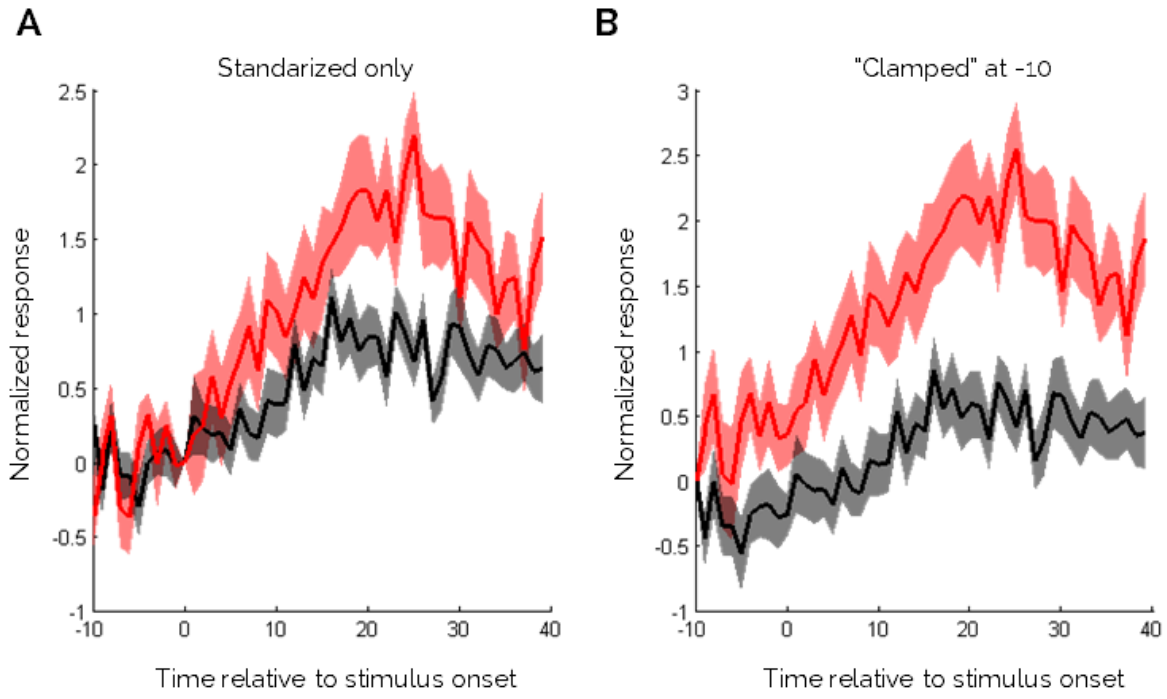

Figure 5.1. Potential artifacts introduced by "clamping" the time course to a single sample prior to stimulus onset. Two simulated time courses averaged across 50 trials with a 2:1 ratio of calm to emotional stimuli are shown. This simulation assumes that the pre-stimulus period is still affected by the signal difference caused by the preceding trial. **A.** Standard baseline correction procedure in which the whole time course was standardized relative to the mean and standard deviation of the entire pre-stimulus period. **B.** The same data but also including the "clamping" procedure in which the response at the first time point is subtracted from the whole time course. This procedure amplifies differences in the pre-stimulus period and can in many cases result in significant differences.

As already mentioned, one way to overcome this problem would be to use a longer baseline period for clamping and ensuring that this is a true baseline (i.e. that the signal has recovered from the previous stimulus). However, this cannot fix another, potentially greater, problem with the clamping method: its arbitrariness. In standard experiments assuming forward causality it is obvious which part of the measurement constitutes the baseline. In presentiment experiments this choice is unclear. Where should we expect the true baseline to fall? Is it at 3 seconds before the stimulus, or at 10 seconds? What about several days, months, or years? The experimenters in these studies are of course limited by the temporal design of their experiment but the point still stands. Since we don't have any inkling as to the mechanism by which presentiment is supposed to occur we cannot really know where the true baseline should be. Naturally, the further back in time we go relative to the stimulus the more likely it is that the signal is also diluted by the responses to other stimuli, both inside and outside the lab (and – if you believe that retrocausality plays any role – stimuli both before and after the signal will dilute it!). It becomes extremely messy.

While Mossbridge and colleagues suggest that the pre-stimulus period is similar in most of the studies they surveyed, there is in fact an enormous variability. The range of pre-stimulus periods reported is 1-12 seconds. Even when limiting this to only electrodermal experiments the range is almost as wide. How were these times chosen? There are a lot of researcher degrees of freedom in this. It would be completely trivial if a researcher moved back the time point for clamping (or even the time window used as baseline) until they found a significant pre-stimulus effect. Please note that I certainly do not want to accuse all the researchers who conducted these studies of this sort of cherry-picking – but considering the variability in baseline periods it is also difficult to rule out that this is what happened in some cases. At the very least, there should be a commonly agreed protocol for these kinds of experiments so that they can be repeated.

6. Potential artifacts introduced by filtering

I did not touch this point in my article. However, another necessary question to ask about these presentiment studies is in how far it is possible that these effects are simply artifacts of signal filtering during pre-processing. The original authors discussed this but dismissed this point very quickly. This issue certainly demands a lot more scrutiny. One way to address this would for example be to take simulated data with similar response patterns of arousing and calm trials but for which we know that there is no significant pre-stimulus difference. These simulated data can be analyzed using the same pipeline as the real data. If this produces significant pre-stimulus differences this will be very strong evidence that the effect is merely an artifact. A subtler variation of this approach may also allow researchers to quantify how large a presentiment effect would have to be not to be caused by a filtering artifact. None of the presentiment studies I read performed such control analyses.

7. Randomization and counterbalancing

One argument for why according to parapsychologists most of the "mainstream" literature is not useful for quantifying these presentiment effects (and why they cannot be included in the meta-analysis even though some researchers have claimed to have found presentiment in "mainstream" data) is that these studies used counterbalanced trial orders. Counterbalancing is when the order of stimuli is not truly random but when each participant is presented with the same number of target and control trials but the order of these trials is randomized (sometimes this is also called "randomization without replacement"). More extreme forms of counterbalancing do not only ensure that the number of trial types is equal for each participant, but they also balance the occurrence of particular trial sequences up to a certain number of trials. So for instance, they would ensure that the sequence of control-control appears as often as the sequence target-target. In really extreme cases they may do this for sequences longer than two trials.

The purpose of such counterbalancing is to *ensure that temporal order effects are minimized*. It is one of the ironies of this whole discussion that parapsychologists regard mainstream results as confounded due to counterbalancing, when counterbalancing would in fact be the optimal approach for controlling the temporal order confounds in presentiment studies.

Another irony about this issue is that parapsychologists are very worried about participants being able to predict the upcoming stimuli in counterbalanced sequences. However, the prerequisite for them to do that is that they have advance knowledge that the sequence is counterbalanced. They can only have this knowledge through precognition or if they have experience with the nature of the experiment, for example if they have participated in several of these experiments in the past (which is a reason not to reuse the same participants in such experiments). Without this, they could not possibly tell apart a counterbalanced from a truly random (probabilistic) sequence, in particular if the number of trials is sufficiently large (Figure 7.1).

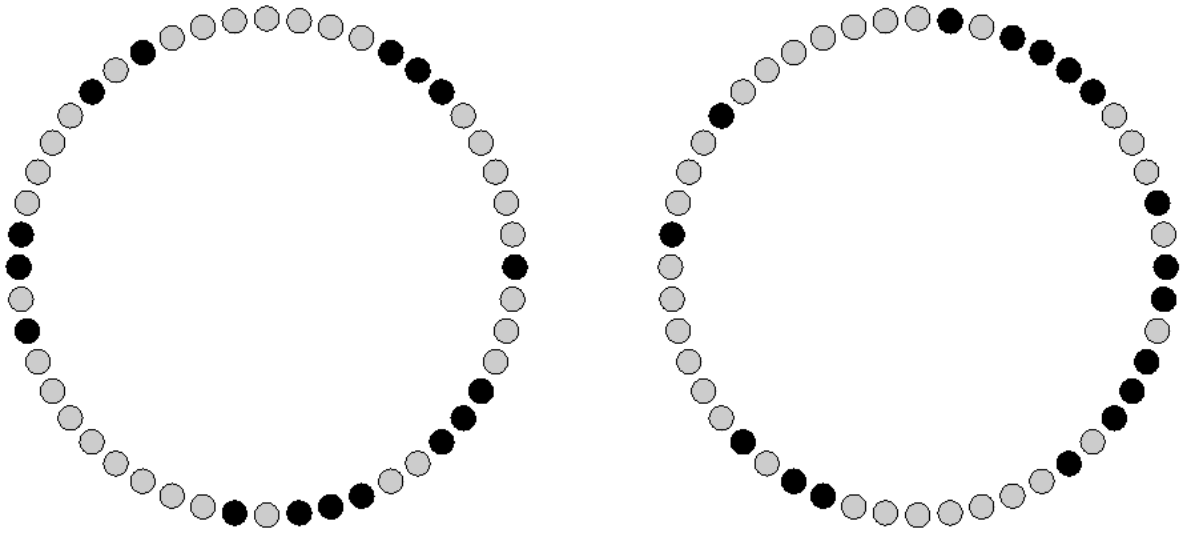

Figure 7.1. Two randomly generated patterns. In one the color of each disk is determined probabilistically. The other has a predefined number of black and gray disks that have been shuffled. Without any knowledge of how many black disks there were, there is no way to tell the patterns apart.

8. Publication bias and the larger literature

While there have been relatively few studies investigating presentiment effects, the paradigm used by these researchers is extremely common. There must be hundred or more studies using emotional and calm stimuli, probably using a similar odd-ball design, which could be included in the meta-analysis. Mossbridge and colleagues performed a publication bias analysis. The most conservative outcome of this is that one would need 87 additional data sets showing no presentiment effect for their results not to support the existence of presentiment. I am willing to bet that there are far more than 87 data sets out there in the literature which could have been included. Obviously they would have to relax the rule about not having counterbalanced randomization, which (as I already discussed) is a straw man argument and would in fact improve the quality of the data by controlling temporal order effects.

Of course, if it was indeed the case that including a large number of "mainstream" studies obliterates the evidence for presentiment, this would raise the question why it is largely parapsychologists who find these effects. I believe that these effects are the result of experimental and analytical artifacts such as the ones I described already. Therefore I would wager that the presentiment effects disappear if these presentiment experiments were replicated with standard methods (counterbalanced orders, no clamping, etc). Conversely, if the mainstream data are analyzed using these flawed methods, I won't be surprised if they also show presentiment although the artifacts are more likely to cancel out with a larger database.

9. How biologically plausible are presentiment effects?

I further questioned the biological plausibility of these presentiment effects. A hypothesis must be consistent in itself. Let's assume that presentiment exists, that is, that there actually are neuronal events prior to stimulus onset that predict it and could be used for participants for conscious precognition. Don't such neural events cause (in the forward direction) metabolic and physiological responses? It is perhaps reasonable that they would trigger an electrodermal response (i.e. increased sweating because the upcoming stimulus is arousing). The electrodermal response peaks after 3-5 seconds. A hemodynamic response peaks after about 6 seconds and then slowly returns back to baseline over the following 20 or so seconds. Electrophysiological measures like EEG on the other hand are much faster.

Something clearly doesn't add up. If the same pre-sentient neural event causes all of these physiological correlates most of them must bleed in well into the post-stimulus period. Alternatively, if these physiological measures neatly appear at clear times before stimulus onset this means that each is caused by a different neural event. This however seems very improbable because each of these neural events should also cause physiological correlates.

A quick inspection of the pre-stimulus periods used in the presentiment studies included in the Mossbridge meta-analysis shows that for all the EEG studies the presentiment effect was measured within the final second prior to stimulus onset. For the electrodermal experiments the time windows have a wider range. Exactly at what time before the stimulus did these effects occur? Under forward causality the neural event less than 1 s prior to the stimulus cannot have produced an electrodermal response several seconds earlier. Or perhaps the authors are suggesting that not only do these electrodermal effects predict the stimulus presentation in the future, but they also predict the neural presentiment effect in the not-quite-so-distant future which in turn predicts the future stimulus? The whole process seems to become pretty complex.

An alternative hypothesis might be that the physiology gears up to the upcoming future in some anticipatory manner. Hemodynamic responses are thought to occur because neuronal activity causes a metabolic demand in neural tissue and therefore the blood flow to the active regions increases after neuronal firing. It is conceivable that blood oxygenation increases in anticipation of a neuronal event. In fact, as some experiments have suggested [9] hemodynamic activity may occur even in the absence of neuronal activity (please note that these findings are highly controversial but let's assume for the sake of this argument that they are correct). The question I would then ask myself in the context of these presentiment effects, however, is why these pre-stimulus effects are so weak. If the hemodynamic apparatus of the brain were truly able to anticipate the upcoming stimulus, it surely would make sense to increase blood flow prior to stimulus onset on the scale that is required for sustaining the upcoming neural activity.

One presentiment researcher actually formulated the hypothesis that physiological responses are mirrored back in time through some form of time symmetry [10]. Moreover, this hypothesis posited that the physiological event is mirrored not around the time of stimulus onset but around the latency of the physiological response. There is however no clear explanation why this should be so nor why the mirrored response is so much weaker than the actual response (it is suggested that this has to do with brain volume but why this should matter is not entirely clear to me). However, as bizarre as it sounds, this idea seems to be the simplest retrocausal explanation for these effects. I doubt however that the predictions this hypothesis makes are supported by a lot of empirical evidence. Moreover, the hypothesis suggests that these presentiment effects only occur when the participant is conscious of the stimulus. Of course, this neglects the altogether simpler explanation that these effects only occur under conscious awareness because participants only have expectations of any upcoming stimuli if they are aware of them.

Taken together, any formulation of biological mechanisms underlying presentiment effects seems to rest on a whole host of assumptions. This leads me to think that in fact any explanation that can do without all these assumptions is far more probable. Any theory based on analytical artifacts, expectation or trial order effects is far more plausible. For one thing, we already have a large body of compelling evidence that these factors exist.

10. Directly testing expectation bias and order effects

In my opinion the most important specific suggestion I make in the commentary for the original authors is that presentiment effects should be tested directly against the possibility that they are caused by expectation bias or simple trial order. I think I already provide sufficient detail there but to recapitulate: the researchers could run an experiment in which participants are exposed to different stimulus sequences. In the simplest case this could be pairs of trials such as target-target, control-control, target-control, and control-target. This will allow them to directly estimate the pre-stimulus activity prior to the second stimulus under each condition. Naturally the order of these double trials should be randomized and counterbalanced and there should be sufficiently long recovery periods between them in order for the signal to decay to baseline.

Similarly, I described a possible experiment for testing for expectation effects. Participants can be presented with a cue that informs them about the probability that the next stimulus will be an emotional target. This allows experimenters to manipulate expectation directly. In the critical condition the cue makes participants expect a calm control, but the next stimulus is actually an arousing target. Will they still show a presentiment effect?

This design should seem obvious. It directly manipulates the possible confound to test how this influences the variable of interest. *This is one of the few experiments that could provide compelling evidence that these pre-stimulus effects truly reverse causality.* However, as far as I can see the parapsychology community has so far shied away from such experiments. All manner of purported confounds are brought forth as to why such an experiment would be inadequate. However, it is nevertheless clear that this is the only direct way to rule out expectation or order confounds. You can worry about possible problems with this later; the strength of the evidence such experiments would provide strongly outweighs the problems.

11. Are single-trial experiments a solution?

In their articles, Mossbridge and colleague suggest that the solution to the problem of expectation bias is to run single-trial experiments. So each participant is only exposed to one stimulus and the pre-stimulus activity is measured. And sure enough, if this stimulus activity predicts the upcoming stimulus class it is difficult to explain this away with any of the concerns I have raised so far. As such, it is not a bad idea to perform such an experiment. However, this idea is not without problems of its own.

First, it is not necessarily free of expectation bias. While it is true that there is no previous stimulus that could shape expectations and cause adaptation effects, a single-trial experiment necessarily changes the experiment from a within-subject to a between-subject design. This means that great care must be taken to ensure that there are no differences in the groups exposed to the two different stimulus classes. Imagine for example that the target group (who is presented with an arousing stimulus) comprises a slightly larger number of individuals with a tendency of anxiety. That alone could produce a subtle difference in their expectation response. Still, with a sufficiently large number of participants such concerns may be alleviated.

However, a much larger problem with this idea is that *single-trial experiments do not test the same effect!* All the presentiment studies analyzed by Mossbridge and colleagues are of a within-subject phenomenon that is repeatable over an extended period of time. In their second article the authors even speculate about possible practical applications of this effect, including the detection of roadside bombs. Such a device would be of little use if it allowed one use only, especially if one couldn't control when to use it. Moreover, due to the weak signals it also means that the presentiment effect would only be exhibited significantly by a sub-group of the sample (large false negative rate) while there will presumably also be a fairly large number of participants in the control group who will show a significant effect (large false positive rate). This makes it unlikely that this particular measure could have any practical use whatsoever.

While showing that presentiment works at the single-trial level would no doubt be an exciting result, this tells us very little about whether within-subject presentiment effects using repeated trials are the same phenomenon. It would not rule out expectation or order effects in those experiments. It would only be evidence that presentiment occurs in the single trial situation.

The idea to conduct this single trial experiment illustrates yet again a point I made in my commentary and which many critics of psi research have made before me [11]. Most of psi research is concerned solely with showing that psi exists, not with explaining how psi works. There is no other reason to do this single-trial experiment. All it can show us at present is that using yet another paradigm there is evidence for presentiment. It does however not even begin to formulate any theory for why these effects exist.

12. The second law of thermodynamics and the arrow of time

Before I go into this point, let me concede I am not a physicist and my knowledge of quantum mechanics is very limited. But I am willing to bet that the overwhelming majority of theoretical physicists will agree with my basic argument here, and greater minds than mine have argued that macroscopic time reversals are impossible [12].

That being said, the second law of thermodynamics states that in an isolated system the entropy never decreases. It is one of the few physical principles that does not seem invariant to the direction of time and has therefore been seen as a reason why in the macroscopic universe time runs forward. However, like any scientific theory it is only a model. The laws of thermodynamics were the best model to explain physical observations for their time. Since then there have of course been further theoretical developments which have greater explanatory power: Einstein's theories of relativity and, later, quantum mechanics.

There appears to be a lot of discussion in theoretical physics about the conditions under which time travel and reversals of causality can occur. Our present understanding assumes that if time reversal is at all possible it only applies to sub-microscopic scales. There may of course be gaps in our understanding (in fact, I'm counting on it). Thus, it is the reasoning of many psi researchers is that quantum effects may cause time reversals in cognitive function. It is impossible to rule out such effects because this is an untestable hypothesis. Presentiment effects or behavioral evidence of precognition do not confirm it. Quantum effects on biology are an untested *post-hoc* speculation for how precognition might work. At present quantum biology remains pure assumption but provides no explanations and thus makes no clear predictions. As with all scientific theories, one should start with the most trivial explanations that have the greatest explanatory power. Until there is a clear formulation of the predictions that quantum theory makes for cognition and physiology, it frankly is pointless to even consider its existence except as a purely theoretical thought experiment.

The laws of thermodynamics generally do a very good job explaining the macroscopic world. Psi effects notwithstanding, as far as biology is concerned they are fully sufficient to explain our observations. It requires fewer assumptions to come up with explanations that do not reverse causality or which require quantum processes. Therefore, the *most parsimonious explanation* for presentiment or precognition is probably one which is *not based on untestable quantum processes*.

13. The process of scientific investigation

Whatever your finding and however strong your statistical inference, researchers must *always* ask more questions. The purpose of scientific research is to ask why a particular result occurred, not trying to prove wrong all the alternatives to your favorite hypothesis. If you take anything from what I wrote, it should be that *to be a scientist is to be a skeptic*. We do not advance our understanding of the cosmos by clinging to some fixed idea but by being prepared to think outside the box, to be as open-minded as possible and to seek out every possible explanation for our observations.

The irony of this statement isn't lost on me: proponents of psi sometimes liken themselves to great scientists like Galileo or Darwin. It is easy to see the appeal of this. There is something nice about being the underdog who fights for revealing the scientific "Truth" in the face of dogmatic opposition. It is also easy to misinterpret the opposition psi research encounters in "mainstream" science as an attempt by the powers-that-be to suppress the truth, much like how centuries ago the church opposed heliocentrism or the theory of evolution by natural selection.

However, this is not what is happening here. I will concede that many mainstream scientists probably have a strong prior against retrocausality. It will take very convincing evidence for them to update their beliefs in favor of such theories. But in turn it seems to me that psi researchers have a *strong prior for the existence of psi effects* but without any real theoretical basis.

What I want to achieve by critically engaging this work is two things: on the one hand, I hope to encourage researchers to apply scientific reasoning to their data. Instead of going on about retrocausality they ought to start with the simplest explanation, the one requiring the smallest number of assumptions and then slowly work their way towards the best explanation. In my opinion it is wrong to jump from the null hypothesis to the least probable one merely because you think you have ruled out the potential confounds you could think of. And there is no way around it – retrocausality is the least probable hypothesis because there is no clear theory that explains how it works and that makes testable predictions to support it beyond showing that it happens.

My second goal for this discussion is to remind all my fellow researchers to remain skeptical including of our own results. As the quote by Richard Feynman in my commentary suggests there is nobody we can fool more easily than ourselves (I must add at this point that the credit for suggesting this quote goes to one of the reviewers of my article, Russ Poldrack). This is something we all too often forget, myself very much included. Just because we find a nicely expected result does not mean our hypothesis is correct. You can always dig deeper. I hope that discussions like this help emphasize the importance of this and that critical self-analysis and scientific skepticism will in future be valued more by fellow scientists and grant committees alike.

References

1. Mossbridge, J. A. *et al.* Predicting the unpredictable: critical analysis and practical implications of predictive anticipatory activity. *Front. Hum. Neurosci.* **8,** 146 (2014).
2. Mossbridge, J., Tressoldi, P. & Utts, J. Predictive physiological anticipation preceding seemingly unpredictable stimuli: a meta-analysis. *Front. Psychol.* **3,** 390 (2012).
3. Bem, D. J. Feeling the future: experimental evidence for anomalous retroactive influences on cognition and affect. *J. Pers. Soc. Psychol.* **100,** 407–425 (2011).
4. Bierman, D. & Scholte, H. A fMRI brain imaging study of presentiment. *J. Int. Soc. Life Inf. Sci.* **20,** 380–389 (2002).
5. Bennett, C. M., Baird, A. A., Miller, M. B. & Wolford, G. L. Neural correlates of interspecies perspective taking in the post-mortem Atlantic Salmon: An argument for multiple comparisons correction. (2009). at <http://prefrontal.org/files/posters/Bennett-Salmon-2009.jpg>
6. Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S. F. & Baker, C. I. Circular analysis in systems neuroscience: the dangers of double dipping. *Nat. Neurosci.* **12,** 535–540 (2009).
7. Vul, E., Harris, C., Winkielman, P. & Pashler, H. Puzzlingly High Correlations in fMRI Studies of Emotion, Personality, and Social Cognition. *Perspect. Psychol. Sci.* **4,** 274 –290 (2009).
8. Radin, D. Electrodermal presentiments of future emotions. *J. Sci. Explor.* **18,** 253–273 (2004).
9. Sirotin, Y. B. & Das, A. Anticipatory haemodynamic signals in sensory cortex not predicted by local neuronal activity. *Nature* **457,** 475–479 (2009).
10. Bierman, D. Consciousness-induced restoration of time symmetry (CIRTS). A psychophysical theoretical perspective. Journal of Parapsychology, 24, 273–300. *J. Parapsychol.* **24,** 273–300 (2010).
11. Alcock, J. E. Give the Null Hypothesis a Chance: Reasons to Remain Doubtful about the Existence of Psi. *J. Conscious. Stud.* **10,** 29–50 (2003).
12. Hawking, S. W. Chronology protection conjecture. *Phys. Rev. D* **46,** 603–611 (1992).