CORTEX 93 (2017) 119–132

Available online at www.sciencedirect.com

ScienceDirect

Journal homepage: www.elsevier.com/locate/cortex

Research report

Transcranial magnetic stimulation to visual cortex induces suboptimal introspection

Megan A.K. Peters [a,*], *Jeremy Fesi* [b], *Namema Amendi* [b], *Jeffrey D. Knotts* [a], *Hakwan Lau* [a,c] *and Tony Ro* [b]

[a] *Department of Psychology, University of California, Los Angeles, Los Angeles, CA, USA*
[b] *Programs in Psychology and Biology, The Graduate Center of the City University of New York, New York, NY, USA*
[c] *Brain Research Institute, University of California, Los Angeles, Los Angeles, CA, USA*

ARTICLE INFO

Article history:
Received 31 January 2017
Reviewed 3 April 2017
Revised 16 April 2017
Accepted 23 May 2017
Action editor Sven Bestmann
Published online 2 June 2017

Keywords:
Bayesian ideal observer
Blindsight
Metacognition
Noninvasive neuromodulation
Unconscious perception

ABSTRACT

Blindsight patients with damage to the visual cortex can discriminate objects but report no conscious visual experience. This provides an intriguing opportunity to allow the study of subjective awareness in isolation from objective performance capacity. However, blindsight is rare, so one promising way to induce the effect in neurologically intact observers is to apply transcranial magnetic stimulation (TMS) to the visual cortex. Here, we used a recently-developed criterion-free method to conclusively rule out an important alternative interpretation of TMS-induced performance without awareness: that TMS-induced blindsight may be just due to conservative reporting biases for conscious perception. Critically, using this criterion-free paradigm we have previously shown that introspective judgments were optimal even under visual masking. However, here under TMS, observers were *suboptimal*, as if they were *metacognitively blind* to the visual disturbances caused by TMS. We argue that metacognitive judgments depend on observers' internal statistical models of their own perceptual systems, and introspective suboptimality arises when external perturbations abruptly make those models invalid – a phenomenon that may also be happening in actual blindsight.

© 2017 Elsevier Ltd. All rights reserved.

Neurological cases of blindsight (Weiskrantz, 1986, 1996) present an intriguing opportunity for studying consciousness (Giles, Lau, & Odegaard, 2016): patients with damage to primary visual cortex can discriminate targets above chance yet report no conscious visual experience of the stimuli (Cowey & Stoerig, 1997, 1991, 1995; Kentridge, Heywood, & Weiskrantz 1999, 2004; Sahraie, Hibbard, Trevethan, Ritchie, & Weiskrantz, 2010). However, such patients are rare and symptoms are often heterogeneous. In response, many researchers have sought to elicit blindsight-like unconscious perception in neurologically intact observers using visual masking or other stimulus manipulations (Breitmeyer, Hoar, Randall, & Conte, 1984; Breitmeyer, 2007; Charles et al., 2016; Charles, King, & Dehaene 2014; Fogelson, Kohler, Miller, Granger, & Tse, 2014; Kolb & Braun, 1995; Ramsøy & Overgaard, 2004; Sandberg, Timmermans, Overgaard, & Cleeremans, 2010).

* *Corresponding author*. 1285 Franz Hall, Box 951563, Los Angeles, CA, USA.
E-mail address: meganakpeters@ucla.edu (M.A.K. Peters).
http://dx.doi.org/10.1016/j.cortex.2017.05.017
0010-9452/© 2017 Elsevier Ltd. All rights reserved.

However, some researchers have pointed out that many of these studies could be contaminated by criterion bias: observers may report 'unseen' only because the stimulus fell below some arbitrary threshold for reporting 'seen', not because the stimulus was truly unconscious (Eriksen, 1960; Hannula, Simons, & Cohen 2005; Lloyd, Abrahamyan, & Harris, 2013; Merikle, Smilek, & Eastwood 2001). Several groups have sought to elicit blindsight-like behavior in normal observers while addressing this confound, but some of these efforts encountered conceptual or replicability problems (Balsdon & Azzopardi, 2015; Evans & Azzopardi, 2007; Kolb & Braun, 1995; Kunimoto, Miller, & Pashler, 2001). Further, it was recently demonstrated that blindsight-like behavior in normal observers does not occur under visual masking conditions once the criterion confound is removed by using a criterion-free task (Peters & Lau, 2015). This suggests that in normal visual masking, criterion bias may indeed be a problem.

Unlike visual masking, transcranial magnetic stimulation (TMS) provides a closer analog to the neuroanatomical deficits exhibited by blindsight patients. Further, it has been demonstrated that TMS to visual cortex results in blindsight-like unconscious perception in normal observers (Boyer, Harrison, & Ro, 2005). Lloyd and colleagues (Lloyd et al., 2013) criticized this study for falling prey to the same criterion bias problem as others, claiming it simply demonstrated near-threshold conscious perception rather than blindsight. We have addressed some of their criticisms elsewhere (Peters, Ro, & Lau, 2016); we also note that other studies from the same lab have shown blindsight-like behavior due to TMS in ways that are less likely to be influenced by criterion biases (Ro, 2008; Ro, Dominique, Lee, & Chang, 2004). Here we sought to empirically examine how TMS-induced changes in subjective visual experience, such as TMS-induced blindsight, may go beyond the effects induced by visual masking alone.

We used a criterion-free 2-interval forced-choice method for subjective ratings (Barthelmé & Mamassian, 2009, 2010; de Gardelle & Mamassian, 2014; Peters & Lau, 2015) to determine whether TMS to visual cortex can induce blindsight-like unconscious perception in normal human observers (Fig. 1A). This task does not require subjects to maintain a response criterion to say 'yes, I saw it' or 'no, I didn't see it' for the subjective rating; instead, observers judge which of two intervals was more visible. It is a conservative test of whether introspective suboptimality can occur due to TMS, since it has been shown that even under visual masking conditions people behave optimally on this task (Peters & Lau, 2015). If TMS-induced "blindsight" is indeed a case of near-threshold conscious perception (Lloyd et al., 2013) and no different from visual masking, we should expect that observers will optimally reduce their visibility ratings in proportion to the reduction in objective discrimination performance caused by disruptions in visual processing due to TMS (Fig. 1B). This is because they have internal knowledge of the statistics governing their sensory inferences (King & Dehaene, 2014; Ko & Lau, 2012; Lau, 2007), including any noise introduced by TMS (Fig. 1C, top row). Alternatively, it has been suggested that observers may be unaware of changes in their sensory processing architecture that can affect perceptual inferences (Seriès, Stocker, & Simoncelli, 2009), which might lead them to introspectively judge noisier samples to be more extreme, leading to higher visibility ratings (Fetsch, Kiani, Newsome, & Shadlen, 2014; Rahnev, Bahdo, de Lange, & Lau, 2012; Rahnev et al., 2011; Rahnev, Maniscalco, Luber, Lau, & Lisanby, 2012; Zylberberg, Fetsch, Shadlen, & Frank, 2016; Zylberberg, Roelfsema, & Sigman, 2014) (Fig. 1C, bottom row).

We used Bayesian observer computational modeling to quantitatively arbitrate between these two hypotheses. Our results indicate that TMS causes introspective suboptimality that can be mechanistically explained by a 'metacognitively blind' Bayesian observer that is "unaware" of the noise that TMS introduces into the visual processing architecture. These results may shed new light on the neurological disorder of blindsight, as well as provide important insight into the mechanisms of higher order metacognitive judgments of perception.

1. Materials and methods

1.1. Subjects

Fourteen subjects (mean age = 28.5; 9 males; 13 right-handed) gave written informed consent to participate in our study. All participants had normal or corrected-to-normal vision. This study was conducted in accordance with the Declaration of Helsinki and approved by the City University of New York's Institutional Review Board.

1.2. Behavioral methods

1.2.1. Stimuli & experimental design

We used a criterion-free two-interval forced choice (2IFC) task to measure visibility assessments in objective decisions (Barthelmé & Mamassian, 2009; de Gardelle & Mamassian, 2014; Peters & Lau, 2015). On each trial, two intervals of an oriented-bar target (subtending .25 × .05 visual degrees) were presented .35 visual degrees to the right of the fixation cross (Fig. 1A). Subjects indicated the orientation of each bar (45° left- or right-tilted from vertical; Type 1 judgment) with keypresses, and then indicated which interval they felt contained the more visible target (Type 2 judgment). The targets were similar to those used by Boyer and colleagues (Boyer et al., 2005), but embedded in the 2IFC task structure. In one interval on each trial, TMS was applied to occipital cortex at one of three latencies ("TMS interval"; see next section for details); in the other interval, no TMS was administered ("noTMS interval").

Stimuli were presented on a Sony Trinitron 17-inch cathode ray tube (CRT) monitor set to a 70 Hz refresh rate, calibrated via gamma correction to make the luminance output profile approximately linear, via MATLAB v. 2012b (Natuck, MA) with PsychToolbox (v. 3.0.10) on an Intel-based Dell computer. Subjects were seated at 57 cm viewing distance from the screen. Orientation discriminations were made via the "<" and ">" buttons, and visibility judgments were made via the "1" and "2" buttons. The target in each interval could take on one of five contrast levels ranging from zero (physically absent) to 100% ("High"). Contrast for the three intermediate contrast levels was titrated to reach 65% ("Low1"),

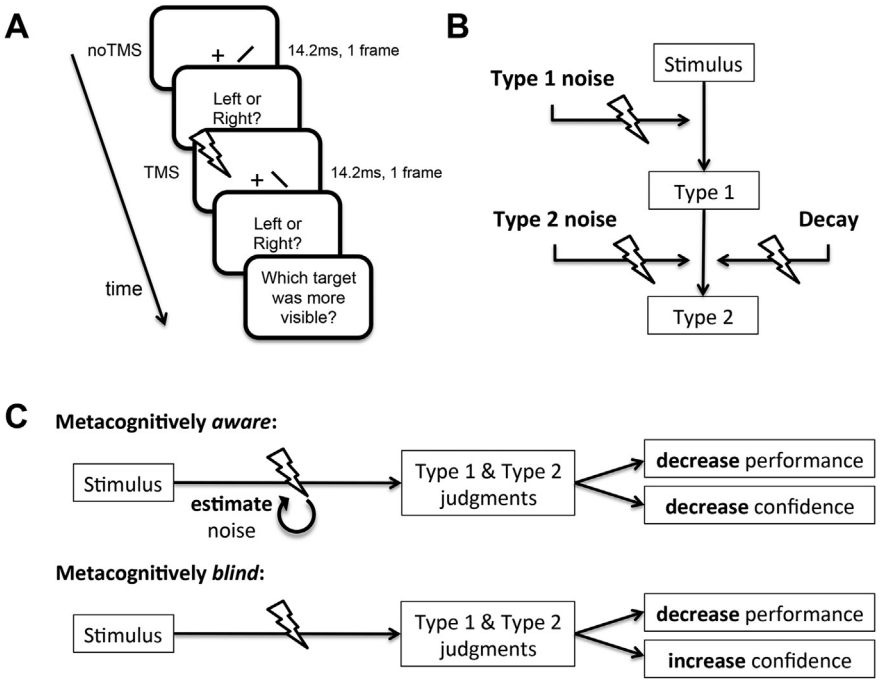

Fig. 1 – Two-interval forced-choice behavioral task and schematic of hypothesized effects of TMS implemented in Bayesian computational models. (A) On each trial, subjects viewed two intervals containing a left- or right-tilted bar target (subtending .25 × .05 visual degrees) at varying contrast levels, presented .35° to the right of the fixation cross. In only one of the intervals (counterbalanced), they received TMS to visual cortex. Subjects indicated whether the target was left- or right-tilted in each interval (Type 1 behavior), and then judged which of the targets was more visible (Type 2 behavior). Because subjects are comparing the visibility between two stimuli that have just been presented, they do not have to maintain an arbitrary criterion for when to report "high" versus "low" visibility. This minimized demand to maintain decision criteria minimizes the effects of Type 2 (subjective) noise on behavior, while removing any response-bias confounds due to criterion setting in the Type 2 judgments. (B) TMS to visual cortex may alter the visual processing stream in several ways, over and above any internal noise or signal decay already present (Maniscalco & Lau, 2016): by adding Type 1 noise (reducing objective performance), adding Type 2 noise (reducing correspondence between accuracy and visibility judgments), and/or decaying the signal (reducing visibility). (C) An ideal observer should be aware of the Type 1 noise being added due to TMS that causes a reduction in performance. In this case, the *metacognitively aware* observer should reduce visibility ratings for stimuli in the TMS interval concomitant with the reduction in objective performance that TMS causes. However, a plausible alternative is that because TMS happens randomly, and because it unnaturally bypasses retinal input, an observer may not be aware of the noise introduced by TMS. This *metacognitively blind* observer will judge the internal visual signal according to the expected statistics of the sensory system, which are less noisy than the TMS-corrupted system. This will lead to an increase in visibility in the TMS interval. Both Type 2 noise and Decay have the effect of reducing visibility in the TMS interval. See main text for details.

75% ("Low2"), and 85% ("Low3") correct performance (respectively) using a Bayesian adaptive staircase implemented by QUEST; during the staircase, subjects only performed left/right discrimination without any indication of which interval contained the more visible target.

In the main experiment, ~10% of trials paired a zero-contrast stimulus with a nonzero-contrast stimulus, and ~2% of trials showed two zero-contrast stimuli across the two intervals. When two intervals with nonzero contrast were shown (~88% of trials), stimuli were more likely to take on one of the three thresholded intermediate contrast levels (~63% of trials paired intermediate–intermediate across Low1, Low2, and Low3; ~12% paired intermediate–High, and ~3% paired High–High). All trials (646 trials total) were presented in counterbalanced pseudorandom order in a full factorial design. Following thresholding, the experiment lasted approximately two-three hours, with breaks to prevent the TMS coil from overheating; subjects were also allowed to take breaks if they became fatigued.

1.3. *TMS methods*

TMS was applied using a Cadwell MES-10 stimulator connected to a 9 cm circular coil. We followed previously-reported procedures for localizing visual cortex (Boyer et al., 2005), initially placing the coil about 2 cm above and 1 cm left of the inion. Visual cortex was functionally localized by having subjects report a 4 digit number that was presented for

14 msec at the center of the monitor while TMS was applied at varying latencies and intensities after the stimulus onset. Visual suppression threshold was defined as the lowest TMS output intensity at an optimal TMS coil position and temporal latency at which subjects were no longer able to report the right two numbers on at least 3 out of 5 trials. The mean intensity of the TMS threshold across subjects was 65% of maximum output, with a range of 45–77%. Once visual cortex was localized and threshold intensity for visual suppression was determined, subjects next performed a simple dot stimulus detection task with the TMS intensity set at 10% above the visual suppression threshold to ensure adequate visual suppression of the stimuli during the experiment.

As in the Boyer et al. (2005) study, TMS was applied at 100, 114, or 128 msec after the onset of an oriented bar in one of the two temporal intervals on each trial. These latencies are ones that have been consistently shown to produce optimal visual suppression (Amassian et al., 1989; Ro, Breitmeyer, Burton, Singhal, & Lane, 2003). Unlike in the Boyer et al. study, however, subjects were not required to report their subjective experience of whether or not they perceived the orientation of the bar, but rather were required to respond to their perceived orientation of the bar in each interval and in which interval they perceived the bar to be more visible, as described above.

1.4. Data analysis

1.4.1. Objective performance

We calculated Type 1 objective discrimination performance as d′ according to signal detection theoretic (SDT) metrics (Green & Swets, 1966; Macmillan & Douglas Creelman, 2004) as $d' = z(HR) - z(FAR)$, where $z(\cdot)$ is the standard z-transform, HR is the hit rate, and FAR is the false alarm rate. To determine the effect of TMS on Type 1 accuracy, for each subject at each contrast level above zero we calculated d′ in noTMS intervals (d'_{noTMS}) and d′ in TMS intervals (d'_{TMS}) for each of the three TMS latencies. We subtract d'_{noTMS} from d'_{TMS} to get a difference score for each contrast level C:

$$\Delta d'_C = d'_{TMS_C} - d'_{noTMS_C} \quad (1)$$

with $C \in$ [High, Low3, Low2, Low1]. For visualization, we binned pairs of [d'_{noTMS_C}, d'_{TMS_C}] in five equally-spaced bins from 0 to 5.5 (Fig. 2A and D).

1.4.2. Absolute blindsight

In neurological cases of blindsight, patients are able to perform a task above chance, yet report no subjective confidence or visual experience of target stimuli [Type 1 blindsight (Brogaard, 2015; Sahraie et al., 2010; Weiskrantz, 1986, 1996)]. To look for *absolute blindsight* effects akin to neurological cases of blindsight and the TMS-induced blindsight effect reported by Boyer et al. (2005), we examined the subset of trials in which contrast on the TMS interval was above zero ($C_{TMS} > 0$) and that in the noTMS interval was zero ($C_{noTMS} = 0$). Examining this subset of trials is akin to asking the question, "When you can discriminate the orientation of the target above chance even though you received TMS, is doing so subjectively different from discriminating *nothing*?" Note that this is a very conservative measure.

For the subset of trials in which the TMS interval contained a nonzero contrast target but the noTMS interval contained a zero contrast target, we calculated the objective performance in the TMS interval for each above-zero contrast level for each subject, averaged across TMS latencies (see Results). Because subjects performed slightly differently at different contrast levels despite thresholding, we binned objective performance as measured by d′ into four evenly-spaced bins ranging from 0 to 5.5. For trials in which contrast in the TMS interval was zero, by definition d′ should be zero (indeed, it is not significantly different from zero; see Results); as any deviations from this expected value can be attributed to noise, we therefore reassigned any non-zero d′ values to zero for this contrast level bin to help highlight the location of the *y*-intercept (Fig. 2, middle column).

The 2IFC task does not provide an absolute metric of Type 2 judgment magnitude due to its criterion-free nature. Instead, the relevant metric is the percent of time the TMS interval is indicated as 'more visible' than that in the noTMS interval ('% more visible'). An ideal observer should indicate the TMS interval's target is more clearly visible whenever it has access to introspective information that orientation discrimination performance in the TMS interval ought to be better than performance in the noTMS interval. For the *absolute blindsight* trials, we calculated the Type 2 '% more visible' in all five Type 1 performance bins.

1.4.3. Relative blindsight

Demonstration of *absolute blindsight* (that discriminating an above-zero contrast target above chance is subjectively no different from discriminating nothing) would indicate that TMS completely abolishes any awareness of the stimulus without abolishing task performance ability. However, it is possible that TMS alters awareness of the stimulus, without abolishing it. This possibility is predicted by previous studies that find that TMS can increase confidence even while reducing task performance (Rahnev, Bahdo, et al., 2012; Rahnev, Maniscalco, et al., 2012).

To check for such a *relative blindsight* effect (Lau & Passingham, 2006), we examined the subset of trials in which a target stimulus was present with above-zero contrast in both TMS and noTMS intervals. (Note: this subset of trials is disjoint from the subset of trials used to examine *absolute blindsight*.) On these trials, *relative blindsight* would be demonstrated if observers' visibility judgments ('% more visible'; see above) differ across the TMS and noTMS intervals despite matched performance (Lau & Passingham, 2006).

For this analysis, we examined the subset of trials in which both intervals contained a nonzero contrast target. Because our task did not explicitly use target conditions in which performance was matched, we calculated a difference score between performance in the TMS and noTMS intervals as above, but this time across all possible combinations of contrast:

$$\Delta d'_{i,j} = d'_{TMS_i} - d'_{noTMS_j} \quad (2)$$

We did this for all possible above-zero contrast combinations of [i, j] $\in$ [High, Low3, Low2, Low1]. When $\Delta d'_{i,j} > 0$, performance in the TMS interval is higher than in the noTMS

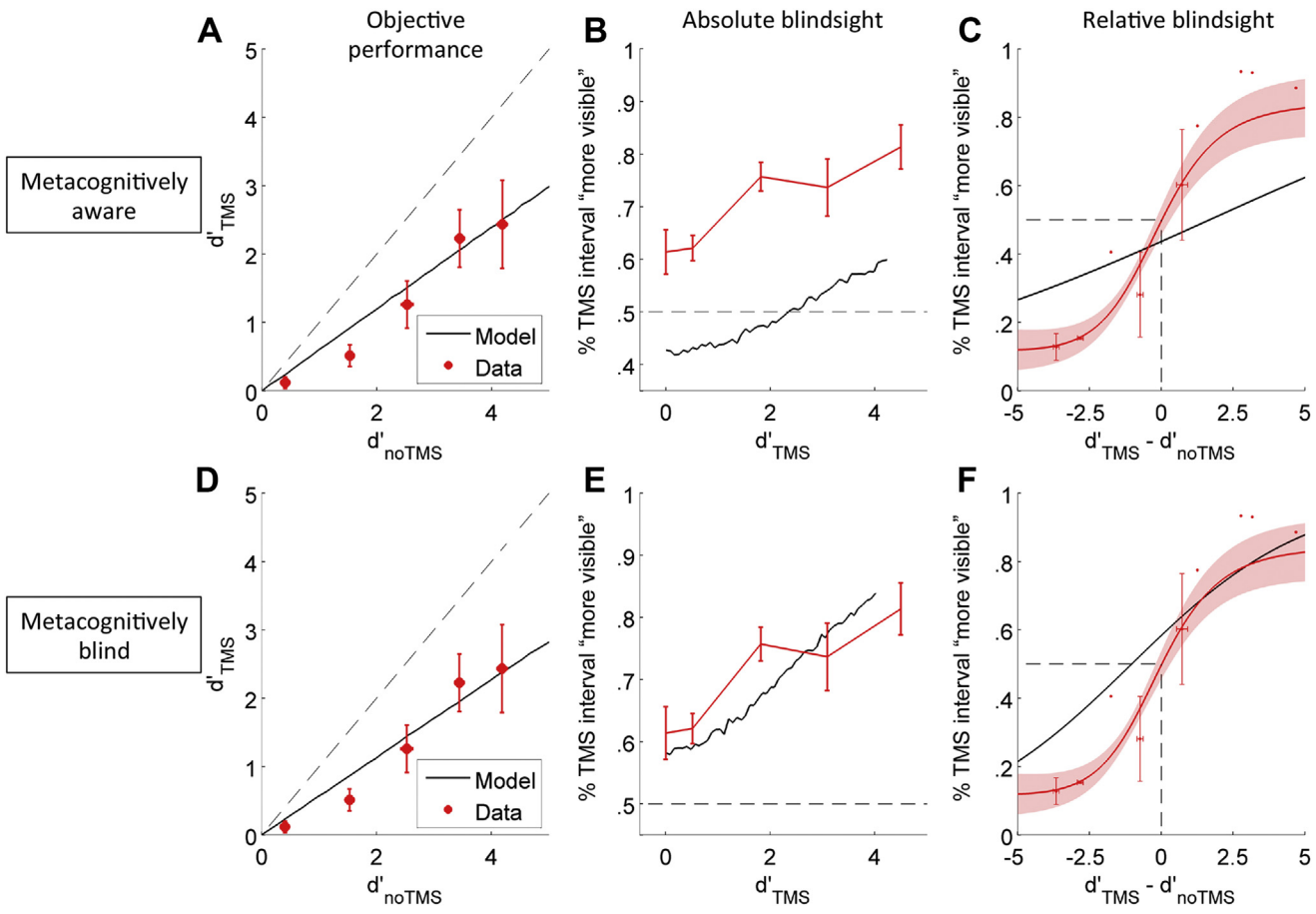

Fig. 2 – Best fitting models that are metacognitively *aware* (panels A–C) and metacognitively *blind* (panels D–F) reveal that both models can predict the Type 1 performance (panels A & D), but the metacognitively *blind* observer provides better fit to the behavioral data at the Type 2 level (panels B, C, E, and F). The metacognitively *aware* observer directly estimates the Type 1 noise introduced by TMS, and so predicts that visibility on the TMS interval – and therefore '% more visible' for the TMS interval – will be reduced in measure to the reduction in objective performance caused by TMS. In contrast, the metacognitively *blind* observer evaluates the noise-corrupted TMS interval samples with reference to the same expected system noise as it uses to evaluate the noTMS interval samples. Because these samples are often extreme due to the TMS noise, the observer judges them to be more confident indicators of the decision it has just made (Fetsch et al., 2014; Rahnev, Bahdo, et al., 2012; Rahnev et al., 2011; Rahnev, Maniscalco, et al., 2012; Zylberberg et al., 2014), and so '% more visible' for the TMS interval is increased. Human subjects showed the same increased '% more visible' scores as the metacognitively *blind* model, especially apparent in panels B and E.

interval, and an optimal observer should indicate that the TMS interval contains the more visible target ('% more visible' > 50%). Likewise, when $\Delta d'_{i,j} < 0$, performance on the noTMS interval is better, and an optimal observer should indicate that the TMS interval does *not* contain the more visible target ('% more visible' < 50%). We fit logistic psychometric functions to the $\Delta d'$ and '% more visible' behavior shown by each subject, of the form

$$y = 1/(1 + 10^{-(ax+b)}) \quad (3)$$

where a is the slope of the psychometric function, and b is the value of the midpoint. We also used these fitted psychometric functions to calculate the '% more visible' at $\Delta d' = 0$ for each subject, which we call the Point of Objective Equality (POE).

1.5. *Computational model*

1.5.1. *Bayesian ideal observer*

Details of the Bayesian ideal observer have been previously described elsewhere (Peters & Lau, 2015). The model assumes that the internal evidence available to an observer on each trial of stimulus strength (a proxy for contrast value) C can be represented as a random sample d drawn from a bivariate Gaussian distribution S_C with $\mu = [C,0]$ (right-tilted) or $\mu = [0,C]$ (left-tilted) and variance Σ, i.e., $d \sim N(\mu,\Sigma)$. Following convention (Barthelmé & Mamassian, 2009; Hedges, Stocker, & Simoncelli, 2011; King & Dehaene, 2014; Knill & Pouget, 2004; Knill & Richards, 1996; Ko & Lau, 2012; Kwon & Knill, 2013; Lau, 2007; Maniscalco, Peters, & Lau, 2016; Peters & Lau,

2015; Stocker and Simoncelli 2006, 2008; Vilares, Howard, Fernandes, Gottfried, & Körding, 2012; Vilares & Körding, 2011; Yuille & Bülthoff, 1996), we assume Σ is a standardized representation of the combination of internal and external noise that the observer has come to expect through experience ($\Sigma = I$, where I is the 2×2 identity matrix), meaning that the observer possesses some knowledge about the statistics of its own perceptual system.

The observer discriminates the target as being right-versus left-tilted by calculating the posterior probability of each according to Bayes rule, marginalized across possible contrast levels C, i.e.,

$$p(S|d) = \int p(S,C|d)dC = \int \frac{p(d|S,C)p(S,C)}{p(d)}dC \quad (4)$$

The observer then judges visibility [or confidence: for the present task, the two can be assumed to produce equivalent behavior (Peters & Lau, 2015)] according to the posterior probability $p(S|d)$ of the discrimination choice it just made, i.e., the probability of having made a correct decision $p(correct)$. It does this for two intervals (two samples d) and then selects the interval with 'clearer visibility', i.e., larger $p(correct)$.

1.5.2. *TMS effect*

Based on previous research, we assumed that TMS to visual cortex may affect internal representations in three possible ways: (1) adding additional *Type 1 noise* to the internal representation of a stimulus, over and above any already-present noise in the system (Rahnev, Maniscalco et al., 2012); (2) increasing *Decay* of the signal between the Type 1 and Type 2 decisions, over and above any already-present Decay (Maniscalco & Lau, 2016); and (3) adding additional *Type 2 noise* to the internal representation after the Type 1 (objective) decision has been made, over and above any already-present Type 2 noise (Maniscalco & Lau, 2016) (Fig. 1B). It is important to note that increasing signal decay prior to a Type 1 judgment has the same effect as decaying the signal between the Type 1 and Type 2 judgments; this is because signal decay does not change the Type 1 decision, only the distance from the decision criterion in SDT terms, and therefore will only affect Type 2 and not Type 1 performance (Maniscalco & Lau, 2016).

To simulate Type 1 noise in the TMS interval, we assume that additive Gaussian noise is added to the sample d drawn on each trial, such that $d^* = d + \varepsilon_1$, where $\varepsilon_1 \sim N([0,0],\sigma_1)$. To simulate Decay, after the Type 1 decision has been made, the noisy internal evidence on a given trial, d^*, is multiplied by a constant ξ, with $0 < \xi < 1$. Thus, $d^{**} = \xi d^*$. Subsequently, $p(S|d^{**})$ is reevaluated as described above to judge confidence/visibility. To simulate Type 2 noise, we add constant Gaussian noise to the posterior probability estimate, such that $p^{**}(S|d^{**}) = p(S|d^{**}) + \varepsilon_2$, where $\varepsilon_2 \sim N(0,\sigma_2)$ (Maniscalco & Lau, 2016). Because probabilities by definition must be between 0 and 1, we also restrict the possible values of $p^{**}(S|d^{**})$ such that $p^{**}(S|d^{**}) = \min(p^{**}(S|d^{**}), 1)$ and $p^{**}(S|d^{**}) = \max(p^{**}(S|d^{**}), 0)$.

1.5.3. *Metacognitively aware versus metacognitively blind*

If TMS-induced "blindsight" is nothing more than near-threshold perception (Lloyd et al., 2013), an ideal observer will *reduce* its confidence/visibility ratings according to the reduction in objective performance caused by TMS Type 1 noise; this is because it is able to update its internal statistical model (Deneve, 2012; Qamar et al., 2013). For example, if TMS reduces percent correct performance from 90% to 70% correct, on average an observer should reduce confidence/visibility ratings from the readout of $p(correct) = 90\%$ to that of $p(correct) = 70\%$. In the 2IFC paradigm used here, this reduction translates to a propensity to indicate that the TMS interval is *less* visible than it otherwise would be, even to the point of judging it to be less visible than the noTMS interval (i.e., '% more visible' < 50%) at low performance levels. This occurs even when the noTMS interval is blank (i.e., in absolute blindsight trials), since the observer does not know *a priori* that the noTMS interval is blank. This reduction in visibility matching reduction in performance implies that the observer possesses perfect knowledge of the Type 1 noise introduced by TMS, i.e., it is *metacognitively aware* of this noise.

This metacognitively *aware* Bayesian ideal observer thus estimates the true covariance structure in the noise-corrupted internal evidence samples, $\Sigma^* = \mathrm{cov}(d^*)$, and makes its orientation discrimination and visibility judgments with this knowledge, e.g., $p(d^*|S) \sim N(\mu,\Sigma^*)$. Because the noise ε_1 is independent of Σ and $\mathrm{var}(X) + \mathrm{var}(Y) = \mathrm{var}(X + Y)$ if $X \perp Y$, the expected value of this covariance is

$$E(\Sigma^*) = \left[k\sigma^2\ 0;\ 0\ k\sigma^2\right] + \Sigma = \left[k\sigma^2 + 1\ 0; 0\ k\sigma^2 + 1\right] \quad (5)$$

Alternatively, it may be possible for an observer to be unaware of noise or other changes in its sensory processing system (Ko & Lau, 2012; Zylberberg et al., 2016; Zylberberg et al., 2014), as has also been suggested in cases of sensory adaptation (Seriès et al., 2009). In the current paradigm, such an observer would be *metacognitively blind*, i.e., unaware that TMS has corrupted its internal representation d via Type 1 noise. The possibility that an observer is metacognitively *blind* to the Type 1 noise introduced by TMS is supported by previous data that demonstrate confidence in perceptual decisions can increase with increasing noise, even as objective performance decreases (Fetsch et al., 2014; Rahnev, Bahdo, et al., 2012; Rahnev et al., 2011; Rahnev, Maniscalco, et al., 2012; Zylberberg et al., 2014). When the observer is metacognitively *blind*, although the internal evidence samples themselves are corrupted by noise, the observer still evaluates them according to the expected noise of the system that it has learned through experience outside the TMS paradigm, i.e., by assuming $p(d^*|S) \sim N(\mu,\Sigma)$. This is not to suggest that metacognitive sensitivity (e.g., meta-d′, Maniscalco & Lau, 2012) is necessarily zero, but instead that the observer makes introspective judgments on the basis of an incorrect internal model. This results in an *increase* in subjective visibility for these samples, as they are judged to be 'extreme' according to the narrower expected noise (Fetsch et al., 2014; Rahnev, Bahdo, et al., 2012; Rahnev et al., 2011; Rahnev, Maniscalco, et al., 2012; Zylberberg et al., 2016; Zylberberg et al., 2014). Thus, for the metacognitively *blind* observer, the predicted effect is an increase in subjective target visibility, manifested as an increase in '% more visible' judgments.

For both the metacognitively *aware* and *blind* observers, Decay and Type 2 noise have the effect of reducing any

extreme visibility value toward *p(correct)* = .5 without affecting Type 1 behavior.

1.5.4. *Parameter estimation and evaluation of model fit*

We fit σ_1, σ_2, and ξ by minimizing the sum of squared error (SSE) between predicted and measured responses at both the Type 1 and Type 2 level across subjects: (1) Decrement in Type 1 d′ (objective performance), (2) Difference between the measured absolute blindsight function and the predicted function, and (3) Difference between the measured psychometric relative blindsight functions and the predicted function. We simulated the expected behavior of each observer for all stimulus strength (i.e., 'contrast') values C ranging from 0 to 5 (corresponding to detection d′ in the 'real world') in steps of .1 using Monte Carlo simulations with 10000 trials at each 'contrast' value. The stimulus strength value that produced the nearest performance to orientation discrimination d′ without TMS was then selected for each subject at each contrast level shown in the experiment, and then the predicted reduction in orientation discrimination d′ and the '% more visible' behavior were observed for that same stimulus strength for both the metacognitively *aware* and metacognitively *blind* models. To seek the best fitting parameter values for each model we pooled all data from all subjects across all conditions and minimized the SSE via a Matlab implementation of the Nelder-Mead Simplex algorithm (fminsearch).

We quantitatively compared the models' fits to the data at each measure of Type 1 and Type 2 behavior by calculating the percent variance explained (R^2) at each. To compare the overall model fits to the data across all behavioral response levels, we compared the SSE at the best fitting values, and also calculated the log-likelihood of the data given the model (*LL*). (Direct comparisons of this kind are warranted because the *aware* and *blind* models have equal complexity, i.e., the same number of free parameters.)

To calculate each model's *LL*, we relied on the formal definition of the likelihood of a certain model *m* with a given set of parameters ϕ:

$$L_m(\phi|data) \propto \Pi_{ij} P_\phi(R_i|S_j)^{n_{data}(R_i|S_j)} \quad (6)$$

where each R_i is a behavioral response a subject may produce on a given trial, and each S_j is a type of stimulus that might be shown on that trial. The expression "$n_{data}(R_i|S_j)$" is a count of how many times a subject actually produced R_i after being shown S_j. The expression "$P_\phi(R_i|S_j)$" denotes the probability with which the subject produces the response R_i after being presented with S_j, according to the model specified with parameters ϕ. This corresponds to the percentage of time each of the models described above produced response R_i after having been "presented" with stimulus S_j. Note that this approach does not examine the performance of a model relative to the behavioral data with reference to any summary statistics, but calculates the model's likelihood with respect to the full distribution of behavioral responses provided by subjects.

1.6. *Control study*

Fifteen subjects (mean age = 29.0; 9 males; 12 right-handed) participated in the control study, which occurred as a pilot to the main study (see next paragraph). Subjects were recruited using the same method as in the main study. All subjects gave written informed consent to participate, and the study was conducted in accordance with the Declaration of Helsinki and approved by the City University of New York's Institutional Review Board. Two subjects were excluded because they did not finish the task (one completed only nine trials, the other only seven trials), leaving 13 subjects in the control group.

All materials, methods, and procedures for this pilot study were identical to the main experiment, with one exception: in this pilot study (which now serves as the control study), a coil holder (Manfrotto Magic Arm, Cassola, Italy) that does not compensate for head movements was used. A disadvantage of using such coil holders is that small head movements typically result in a slight mismatch between the targeted visual suppression area and the actual area stimulated by TMS, which are enough to reduce or eliminate visual suppression. This pilot study therefore allowed us to evaluate whether manually compensating for head movements would be necessary to induce successful suppression in the main experiment. Because all procedures – thresholding, task procedure, and TMS application – were otherwise identical, these pilot data provide an ideal control for the main experiment, in which even the experimenter (a junior research assistant) was unaware that the procedure might not induce optimal visual suppression; in this way, the control study is in fact double-blind controlled.

2. Results

2.1. *Objective performance*

We first examined participants' ability to judge the orientation of the tilted line target as being left or right tilted from vertical. As expected, when contrast was zero in the TMS interval, performance was not significantly different from chance (d′ = 0) [t(13) = 1.83, $p > .05$], and the same was found for zero contrast in the noTMS interval [t(13) = 1.74, $p > .05$]. Thresholding procedures (see Methods) were successful at titrating performance for the three lower contrast levels without TMS, and performance was highest for High (100% contrast) as expected: $d'_{High} = 3.11 \pm 1.24$ (% correct = 89.5% ± 11.3%); $d'_{Low3} = 2.54 \pm 1.37$ (% correct = 84.6% ± 16.2%); $d'_{Low2} = 1.82 \pm 1.09$ (% correct = 77.9% ± 14.6%); $d'_{Low1} = .86 \pm .83$ (% correct = 64.0% ± 13.0%).

We evaluated the degree to which TMS caused a deficit in objective performance by subtracting the performance in the noTMS interval from performance in the TMS interval for all contrast levels above zero (Eq. 2; see Methods). As expected, TMS significantly reduced Type 1 performance (mean $\Delta d'_C = -1.11$), and d′ increases as expected with contrast [2 (TMS: on/off) × 4 (contrast) × 3 (TMS latency) repeated measures analysis of variance (ANOVA): main effect of TMS, $F(1,13) = 16.863$, $p = .001$; main effect of contrast, $F(3,39) = 21.668$, $p < .001$; no other main effects or interactions were significant] (Fig. 2A and D). Because the reduction in d′ was not a function of TMS latency, consistent with previous reports (Boyer et al., 2005), for all subsequent behavioral and modeling analyses we collapsed across TMS latency.

With best-fitting values for Type 1 noise (σ_1), Type 2 noise (σ_2), and Decay (ξ) (Table 1), both the metacognitively *aware* and metacognitively *blind* observers predict a similar decrement in objective performance (*aware*: $\Delta\ d'_C = -1.35$, or mean reduction of 40%; *blind*: $\Delta\ d'_C = -1.45$, or mean reduction of 45%; Fig. 2A and D). Likewise, when only examining Type 1 performance, the *aware* and *blind* models fit the data very well and nearly equivalently, with $R^2_{Type1\ aware} = .926$ and $R^2_{Type1\ blind} = .933$.

2.2. Metacognitive behavior

The similarity between the *aware* and *blind* models observed in the Type 1 behavior abruptly diverges at the Type 2 behavioral level.

2.2.1. Absolute blindsight

In the *absolute blindsight* trials (TMS interval contrast above zero, $C_{TMS} > 0$; noTMS interval contrast at zero, $C_{noTMS} = 0$), subjects tended to indicate that the TMS interval's target was *more* visible than the noTMS interval's target even at low contrast levels in the TMS interval. This is consistent with previous reports of increased confidence due to the introduction of noise into the system (Fetsch et al., 2014; Rahnev, Bahdo, et al., 2012; Rahnev et al., 2011; Rahnev, Maniscalco, et al., 2012; Zylberberg et al., 2014) (Fig. 2B and E). In fact, this '% more visible' measure at $d'_{TMS} = d'_{noTMS} = 0$ (which occurs on trials in which both the TMS and noTMS intervals have zero contrast) is significantly above 50% [mean = .614, $t(13) = 2.211$, $p = .046$]; this significantly deviates from what would be predicted if TMS caused no effect at all, or no change at the Type 2 level regardless of any decrement in performance (Peters & Lau, 2015).

With best-fitting parameters, the *blind* model predicts a similar upward shift to that shown by subjects, with predicted '% more visible' at $d' = 0$ of 58.2%, which is not significantly different from subjects' behavior [$t(13) = .618$, $p = .547$] (Fig. 2D). In contrast, the *aware* model predicts a *downward* shift, with predicted '% more visible' at $d' = 0$ of 42.7%, which is significantly smaller than the '% more visible' at $d' = 0$ shown by subjects [$t(13) = 3.629$, $p = .003$] (Fig. 2B). Thus, the *blind* model correctly predicted visibility would increase as a result of TMS in keeping with other findings in the literature (Fetsch et al., 2014; Rahnev, Bahdo, et al., 2012; Rahnev et al., 2011; Rahnev, Maniscalco, et al., 2012; Zylberberg et al., 2016; Zylberberg et al., 2014), but the *aware* model incorrectly predicted visibility would decrease in concert with the reduction in objective performance. The aware model's incorrect interpretation is in line with the hypothesis that TMS simply produces near-threshold conscious perception by reducing introspective reports in concert with a reduction in objective performance [e.g., Lloyd et al. (2013)]. Visual inspection is confirmed by the goodness of fit tests, which find $R^2_{Abs\ blind} = .648$, and $R^2_{Abs\ aware} = -6.692$ – meaning the *aware* model fits substantially *worse* than a horizontal line through the mean of the data.

2.2.2. Relative blindsight

We evaluated whether TMS caused a shift in visibility in the *relative blindsight* trials (in which both intervals contained a nonzero contrast target) by plotting '% more visible' as a function of the difference in objective performance (d′) between the TMS and noTMS intervals (Fig. 2C and F; see Methods). This allows us to examine whether at equal performance across both the TMS and noTMS intervals (i.e., the difference in objective performance is zero), subjects experienced a difference in visibility due to TMS; this would manifest as a 'POE' different from 50% in the '% more visible' behavioral measure (see Methods).

For the *relative blindsight* trials, with best fitting parameters the *aware* observer predicts POE = 43.6% and the *blind* observer predicts POE = 58.3%. Neither of these is significantly different from the POE exhibited by human observers [behavior: POE = 51.7%; *aware*: $t(13) = 1.117$, $p = .284$; *blind*: $t(13) = .915$, $p = .377$]. However, visual inspection alone reveals that the *blind* model fits the data much better than the *aware* model (Fig. 2C and F), which is confirmed by the goodness of fit metrics: $R^2_{Rel\ blind} = .801$, and $R^2_{Rel\ aware} = .583$.

2.3. Model fits and overall comparison

As with the behavioral fitting, similarities in fitted parameter values can be seen between the *aware* and *blind* models at the Type 1 level, but that is where the similarities end (Table 1). In order to try to fit the data, the *aware* model must estimate a very large amount of Type 2 noise ($\sigma_2 = .287$), whereas the *blind* model estimates a minimal amount of Type 2 noise ($\sigma_2 = .060$). The fitted values for Type 2 noise in the *blind* model are much more realistic when considering the behavioral task used in this study: the 2IFC subjective ratings paradigm minimizes the effect of Type 2 noise, and so we should expect that there is little Type 2 noise present in the behavioral data. Both models predict a relatively similar level of Decay.

In sum, it appears the Type 1 noise is responsible for much of the fit to the behavioral data for the *blind* model, but when metacognitive *awareness* of that noise is assumed, the predicted behavioral effect qualitatively and quantitatively diverges from the actual behavioral reports from subjects. The metrics of overall model fit confirm this view, with all three quantitative metrics – mean SSE, mean R^2, and *LL* – indicating that the *blind* model outperforms the *aware* model in predicting humans' behavior on this task (see Methods).

Table 1 – Best fitting parameter values for the metacognitively *aware* and metacognitively *blind* models, and the quantitative metrics indexing their relative degree of fit to the data.

	Type 1 noise (σ_1)	Type 2 noise (σ_2)	Decay (ξ)	mean SSE	mean R^2	*LL*
Aware	1.340	.287	.947	10.702	−1.728	−481.812
Blind	1.452	.060	.790	5.113	.794	−478.513

2.4. Metacognitive semi-awareness?

We also considered the possibility that an observer may be metacognitively *semi-aware*, i.e., that the model "knows" TMS has corrupted its internal representation – but not by how much. We evaluated the degree of metacognitive blindness this metacognitively *semi-aware* observer might have by computing $\Sigma_N = \Sigma + \alpha(\Sigma^* - \Sigma)$ and fitting α to participants' data. In other words, the model may over- or underestimate the noise caused by TMS. This *semi-aware* observer evaluates its internal evidence samples according to $p(d^*|S) \sim N(\mu, \Sigma_N)$.

The metacognitively *semi-aware* model resulted in an average $LL = -477.623$, which is almost equivalent to the metacognitively blind model LL (−478.513) (Eq. 6). However, the *semi-aware* model's higher level of complexity means it is prone to overfitting. We therefore conducted formal model comparisons via the information theoretic measure Bayesian Information Criterion (BIC), which provides a means for comparing models based on their maximum likelihoods while correcting for model complexity. BIC is computed as:

$$BIC = -2 \cdot n(LL) + k \cdot ln(n) \quad (7)$$

where k is the number of free parameters in the model and n is the number of observations (data points) fitted. Lower BIC values are desirable because they indicate higher model likelihood and/or lower model complexity (fewer parameters).

The metacognitively *semi-aware* model's mean BIC was higher than the blind model's BIC ($BIC_{semi\text{-}aware} = 981.01$; $BIC_{blind} = 976.35$), indicating that the additional parameter did not provide a significantly better fit. When examining the fitted parameter value for α, the degree of semi-awareness, the reason for this becomes clear: the best fitting value for α was .0409, meaning that the *semi-aware* model predicts that observers had almost no awareness of the TMS Type 1 noise at all, making it effectively the same as the *blind* model.

2.5. Control study

One concern is that perhaps rather than experiencing subjective inflation of visibility due to TMS, participants responded based on some sort of post-hoc cognitive reasoning or bias, for example, "I felt the 'zap' in that interval, but I am pretty sure I saw nothing in the other interval, so I should probably say the zapped interval was in fact more visible." If it exists, this response bias effect would be most salient in the *absolute blindsight* trials, succinctly captured by the upward shift of the *absolute blindsight* function y-intercept, i.e., the '% more visible' > 50% observed for trials in which *both* intervals contained a 0% contrast target (Fig. 2B and E).

To ensure that the results of the main experiment are not due to such irrelevant top-down or response bias effects, we utilized previously-collected pilot data as a control in which the TMS produced little to no suppression. With such data, we can determine whether subjects bet on the TMS interval significantly more or less often even when no suppression occurred. If we observed any such 'subjective inflation' in the behavioral responses – i.e., '% more visible' > 50% at the y-intercept of the *absolute blindsight* trials when both TMS and noTMS targets had zero contrast – even without visual suppression, we could attribute the results of the main study to these irrelevant top-down factors. In contrast, if this y-intercept does not differ from 50%, then there ought to be little to no top-down effect of feeling the TMS occurring, and subjects' behavior will match that of an ideal observer that has no signal processing noise due to TMS, as was found previously using visual masking (Peters & Lau, 2015). It is not expected that the y-intercept would be lower than 50% in the context of no visual suppression, as the metacognitively aware model selects the TMS interval as 'more visible' less than 50% of the time when both intervals contain a zero contrast target only because it 'knows' that TMS has introduced noise; if TMS introduces no visual processing noise, no downward deviation from 50% is expected even for a metacognitively aware observer.

2.5.1. Control study results

In contrast to the main study, in the control study TMS did *not* significantly reduce Type 1 performance: an omnibus mixed-design ANOVA including all subjects from both experiments with between-subjects factor group (Active/Control) and within-subject factors TMS (on/off), contrast (4 levels), and TMS latency (3 levels) revealed an interaction between Active/Control group and TMS on/off [$F(1,25) = 9.438$, $p = .005$] despite no main effect of Active/Control group [$F(1,25) = 1.032$, $p = .319$]. This means that subjects in the main study and the control study were likely equally good at completing the discrimination task in the noTMS intervals, but subjects in the control study showed no performance deficit due to TMS unlike subjects in the main experiment. Step-down ANOVAs within noTMS and TMS intervals confirmed no main effect of Active/Control group in the noTMS interval [$F(1,25) = .050$, $p = .824$] but a main effect of group in the TMS interval [$F(1,25) = 4.502$, $p = .044$]. This pattern of results demonstrates that a more reliable maintenance of coil position over visual cortex resulted in more successful suppression in the main experiment, but its absence led to less successful visual suppression in the pilot control study.

As any other main effects and interactions for all subjects would be difficult or impossible to interpret given this group interaction, we conducted a step-down within-subjects ANOVA for the control subjects only, akin to the ANOVA conducted for subjects in the main experiment (see Methods). As expected, this 2 (TMS: on/off) x 4 (contrast) x 3 (TMS latency) repeated measures ANOVA revealed no main effect of TMS [$F(1,12) = .851$, $p > .05$], but still the expected main effect of contrast [$F(3,36) = 35.031$, $p < .001$]. Confirming the above group comparison result, we observed an additional main effect of TMS latency [$F(2,24) = 5.101$, $p = .014$] and an interaction between TMS latency and TMS on/off [$F(2,24) = 5.101$, $p = .014$], suggesting that shorter TMS latencies may have produced more suppression in the control study even if overall there was no suppression on average. No other interactions were observed.

To clarify the interaction between TMS latency and TMS on/off, we conducted three separate *t*-tests against zero on the control study data, one for each TMS latency, to determine whether TMS suppression was confined to one latency when it occurred. These tests revealed that only the shortest TMS latency systematically induced suppression: $t_{100\ msec}(12) = 2.195$,

$p = .049$, $t_{114\ msec}(12) = .827$, $p > .05$, $t_{128\ msec}(12) = .042$, $p > .05$. This is consistent with previous findings that show peak suppression reliably occurs at approximately 100 msec and decreases with longer latencies (Amassian et al., 1989; Kammer, Puls, Strasburger, Jeremy Hill, & Wichmann, 2005). Although the 100 msec latency d′ reduction would not survive stringent correction for multiple comparisons, the aim of using the pilot control study data is to examine the effect on '% more visible' behavior when no suppression is present. Therefore, to ensure that the results of this control analysis were not contaminated by possible suppression we discarded the 100 msec latency data for all control subjects.

Finally, to determine whether the '% more visible' when both the TMS and noTMS intervals contain a zero contrast target was significantly higher (or lower) than 50%, we collapsed across the two longer TMS latencies determined to produce no suppression (114 msec and 128 msec). As expected, the mean '% more visible' for these two latencies was not significantly different from 50% when no visual suppression occurred: a conservative two-tailed *t*-test revealed $t(12) = .5729$, $p > .05$, suggesting that when no suppression is present, subjects behave expectedly as ideal observers, selecting the TMS and noTMS intervals equally as 'more visible' when neither interval contains a visible target (Peters & Lau, 2015). The results of this control study therefore indicate that any bias or top-down decision-level effects of the TMS are not strong enough to produce the suboptimal introspection effect observed in the main results.

3. Discussion

Here we used a criterion-free task to show that TMS to visual cortex induced suboptimal introspection by artificially inflating visibility judgments, even as objective performance was impaired. Our findings are congruent with other reports in the literature that increased noise via stimulus (Zylberberg et al., 2016; Zylberberg et al., 2014) or attentional (Rahnev et al., 2011) manipulations, TMS (Rahnev, Maniscalco, et al., 2012), spontaneous neural fluctuations (Rahnev, Bahdo, et al., 2012), or microstimulation (Fetsch et al., 2014) lead to increased confidence in perceptual decisions despite a detrimental effect (or no effect) on performance. This increased subjective report magnitude is thought to occur because an observer's system evaluates extreme signals (due to noise) with respect to the statistics of internal noise it has come to expect via experience in the environment (Lau, 2007). By utilizing the bias-free 2IFC task, we have extended these findings to show that increased visibility ratings, and the resultant suboptimal introspection, are likely not solely due to Type 2 noise effects or criterion bias.

Our findings contrast with a recently-hypothesized account of TMS-induced blindsight as being no more than a case of normal near-threshold conscious perception (Lloyd et al., 2013). If observers reduced their visibility ratings as a result of the decrement in objective discrimination performance caused by TMS, as implemented by our Bayesian *metacognitively aware* ideal observer, then they would have shown less selection of the TMS interval as 'more visible' in proportion to the reduction in objective performance (decrement of ~40%). In contrast, subjects' behavior more closely matched that of the Bayesian *metacognitively blind* suboptimal observer. This is distinctly different from the behavior that has been shown to occur in this paradigm under visual masking (Peters & Lau, 2015): in a previous study, we used forward-backward masking to render low-contrast targets harder to see in an attempt to induce blindsight-like behavior using the same 2IFC procedure employed in the current study, but found that human observers' metacognitive performance matched that of an optimal Bayesian ideal observer. TMS, in contrast, resulted in distinctly *suboptimal* metacognitive behavior, highlighting the difference between visual masking and noninvasive brain stimulation in manipulating introspective reports. The present result is also consistent with the finding that neural encoding of probabilistic information appears to be blind to adaptation, which also alters the neural response to an identical external stimulus (Seriès et al., 2009).

It is true that neither model produced a perfect fit to the behavioral data, because the models are intentionally simple, as constrained by the level of richness afforded by the present data. In particular, one may observe that the *relative blindsight* model predictions appear almost linear, while human subjects' behavior is more sigmoidal (Fig. 2C and F). This appearance of linearity arises from the relatively large amount of Type 2 noise required to fit even the *metacognitively blind* model, indicating that our Type 2 noise parameter may be absorbing other sources of noise. It is also possible that other more complex models might have produced better fit by considering sources of contributions to the metacognitive signal that do not arise strictly from the feed-forward model architecture shared by both the *aware* and *blind* models. For example, it has been suggested that areas involved in motor planning or execution may also contribute to metacognitive computations (Fleming & Daw, 2017; Fleming et al., 2015). Unfortunately, our current study design precluded investigation of this possibility, as such non-sensory factors were not manipulated and therefore any parameter added to the models to capture such effects would be conflated with existing model parameters, making the models underconstrained. Future studies should combine these approaches to more comprehensively measure the influence of sensory versus non-sensory information on metacognitive judgments.

Despite quantitatively imperfect fits, from qualitative inspection alone it appears the critical factor is whether the observer is aware of the noise introduced by TMS. Future studies should also match external noise in the stimulus to the observed d′ deficit while making observers fully aware of the stimulus manipulations; if observers are successfully made metacognitively aware of the unreliability of the stimulus, they should exhibit the near-threshold behavior hypothesized by Lloyd et al. (2013), consistent with the metacognitively aware Bayesian observer here.

One possible question is why the *metacognitively aware* Bayesian observer did not predict 50% '% more visible' judgments for the TMS interval when objective performance in the TMS interval was at chance ($d' = 0$). In these trials, both intervals had zero contrast, and one might expect that an ideal observer would therefore judge the TMS and noTMS intervals' targets to be about equally visible, leading to 50% '% more visible' judgments. However, the *metacognitively aware*

observer is not privy to the information that both the TMS and noTMS intervals actually have zero contrast. The observer only has access to the internal evidence on every trial, and its knowledge about the reliability of that internal evidence as a result of TMS or noTMS. Because the observer is *aware* that TMS intervals have lower reliability (higher noise), the observer judges visibility in the TMS interval to be less than in the noTMS interval, and so indicates it is 'more visible' less than 50% of the time.

It is also important to address concerns about cognitive biases in the Type 2 judgments. Perhaps subjects felt the TMS and engaged in some sort of top-town reasoning especially in trials where the target was zero contrast in both intervals, i.e., the y-intercept in *absolute blindsight* trials (Fig. 2A and D). This reasoning could be summarized as something like, "I felt the TMS in this interval, but I am pretty sure I saw nothing in the other interval, so I should probably say the TMS interval was more visible." However, our control study demonstrated that when TMS produced no visual suppression, there was also no subjective inflation of '% more visible' behavioral responses: subjects indicated the TMS and noTMS intervals were 'more visible' equally when the TMS did not interfere with visual processing, behaving as ideal observers (Peters & Lau, 2015). This result shows that top-down cognitive biases could not have produced the suboptimal subjective inflation behavior observed in the main experiment.

It is also possible that the TMS pulse caused visual phosphenes, and perhaps subjects based their visibility judgments ('more visible' interval selection) on the *phosphenes'* 'visibility' rather than the targets' visibility, leading to the inflated visibility subjects reported for TMS intervals. However, this would mean subjects are reporting visibility of something other than the target, which would also imply that TMS may have caused a deficit in the ability to distinguish between the visibility of the target and the 'visibility' of a phosphene; such a deficit would also imply suboptimal introspection, albeit of a slightly different variety than is concluded here. This is a common issue in TMS studies, but we do think it is unlikely to adversely affect our results because the TMS parameters in our study were set based on suppression thresholds rather than phosphene thresholds, meaning that there were likely very few trials in which this confusion may have come into play. However, to conclusively rule this possibility out, future studies should examine whether 'visibility' judgments are more dependent on phosphene reports than assumed here.

It has been suggested that subjective confidence and subjective visibility ratings are not equivalent, as they are used here (Overgaard & Sandberg, 2012; Sandberg et al., 2010). Yet they do share important similarities (Fleming & Dolan, 2012; Fleming, Dolan, & Frith, 2012; King & Dehaene, 2014), and in the 2IFC criterion-free subjective rating task and Bayesian computational models employed here, the two kinds of subjective ratings have shown to produce highly similar behavior (Peters & Lau, 2015). However, it has also been reported that relative blindsight (Lau & Passingham, 2006) induced by metacontrast masking may occur with visibility ratings but not confidence judgments (Maniscalco & Lau, 2016). We therefore elected to use visibility ratings, on the assumption that if anything, doing so may give us a better chance of revealing the effect of TMS on subjective judgments.

To the extent that the current results speak also to mechanisms for confidence, our findings are compatible with the view that confidence is represented at a later stage of processing that occurs separately from Type 1 computations (Maniscalco & Lau, 2016). Others have shown that arousal may be a contributing factor in dissociations between confidence and objective performance capacity (Allen et al., 2016). Though it may be argued that TMS may have caused heightened arousal, our control study suggests that this cannot be the explanation of the behavioral effects. In the same work the authors also showed that 'variance' reduces confidence, but there 'variance' refers to something rather different. It concerns the angular spread of a motion signal, rather than the typical kind of random noise supposedly induced by random-dot kinematogram or brain stimulation, or the kind of trial-by-trial variability of signal characterized by 'variance' terms in SDT models (Fetsch et al., 2014; Rahnev, Bahdo, et al., 2012; Rahnev et al., 2011; Rahnev, Maniscalco, et al., 2012; Zylberberg et al., 2016; Zylberberg et al., 2014). Though the mechanisms may differ between these studies, the important message is that there are a number of convincing demonstrations where confidence and objective performance capacity can dissociate.

It appears that the dissociation between introspection and actual sensitivity shown here happen precisely because the observer is not aware of the noise introduced by TMS. This observation may shed light on an important question regarding neurological cases of blindsight: how could a deficit at the objective processing level produce suboptimal subjective experience? The answer may be that an observer comes to expect, through experience, that its internal signal processing environment will exhibit certain statistical properties (Lau, 2007); this assumption is the foundation of Bayesian descriptions of perceptual decision-making (Barthelmé & Mamassian, 2009; Hedges et al., 2011; Knill & Pouget, 2004; Knill & Richards, 1996; Kwon & Knill, 2013; Stocker and Simoncelli 2006, 2008; Vilares et al., 2012; Vilares & Körding, 2011; Yuille & Bülthoff, 1996). When the expected statistical structure is abruptly violated, however, the observer may experience a strong deviation from expected sensory precision and be unable to properly update its metacognitive evaluation computations (Seriès et al., 2009; Zylberberg et al., 2016). This view has been put forth as a potential explanation for neurological cases of blindsight (Ko & Lau, 2012), linking metacognitive computations with subjective awareness.

One might ask, therefore, why in neurological cases of blindsight, the signal processing architecture does not *eventually* update to reflect the new statistical properties of the system after years of subsequent experience. One possible explanation is that the damage to visual cortex, which caused the disorder itself, may preclude the effective updating of a metacognitive representation of the system's statistical properties. How to learn the statistics of the signal processing architecture is not trivial, nor is how such statistics are stored or represented (Lau, 2007). For example, this information might be stored in the connectivity between sensory areas and higher order regions such as prefrontal cortex (Fleming & Dolan, 2012; Fleming, Huijgen, & Dolan, 2012; Lau & Rosenthal, 2011; Rounis, Maniscalco, Rothwell, Passingham, & Lau, 2010), so that perhaps after V1 is damaged these connections may not be able to recalibrate. Indeed, it has been

suggested that drastic changes in signal processing cause even simple models of metacognitive behavior to fail to update their confidence criteria (Ko & Lau, 2012).

Neurological cases of blindsight are typically thought to manifest as above-chance objective performance capacity in the absence of subjective awareness of the stimulus. However, it has also been suggested that subjective experience of visual stimuli can exist in blindsight patients in the absence of visual phenomenology, or qualia – sometimes called Type 2 blindsight (Brogaard, 2015; Cowey & Stoerig, 1995; Foley, 2015; Foley & Kentridge, 2015; Kentridge, 2015; Sahraie et al., 2010; Weiskrantz, 1986, 1996). It is unknown how blindsight patients would perform on a criterion-free subjective task such as the present 2IFC paradigm, in which saying "I see nothing" or "I experience nothing" is not an option. Future studies should explore whether blindsight patients may perform similarly to the TMS-induced suboptimal introspection we report here.

Funding

This work is supported by funding from the Templeton Foundation (grant 21569 to H.L.), the US National Institute of Neurological Disorders and Stroke (NIH R01 NS088628 to H.L.), and the National Science Foundation (NSF EFRI 1137172 to T.R., NSF BCS 1358893/1561518 to T.R).

REFERENCES

Allen, M., Frank, D., Samuel Schwarzkopf, D., Fardo, F., Winston, J. S., Hauser, T. U., et al. (2016). Unexpected arousal modulates the influence of sensory noise on confidence. *eLife*, 5(October), e18103. eLife Sciences Publications Limited.

Amassian, V. E., Cracco, R. Q., Maccabee, P. J., Cracco, J. B., Rudell, A., & Eberle, L. (1989). Suppression of visual perception by magnetic coil stimulation of human occipital cortex. *Electroencephalography and Clinical Neurophysiology*, 74(6), 458–462.

Balsdon, T., & Azzopardi, P. (2015). Absolute and relative blindsight. *Consciousness and Cognition*, 32(October), 79–91. Elsevier Inc.

Barthelmé, S., & Mamassian, P. (2009). Evaluation of objective uncertainty in the visual system. *PLoS Computational Biology*, 5(9), e1000504. Public Library of Science.

Barthelmé, S., & Mamassian, P. (2010). Flexible mechanisms underlie the evaluation of visual confidence. *Proceedings of the National Academy of Sciences*. http://dx.doi.org/10.1073/pnas.1007704107/-/DCSupplemental. www.pnas.org/cgi/doi/10.1073/pnas.1007704107.

Boyer, J. L., Harrison, S., & Ro, T. (2005). Unconscious processing of orientation and color without primary visual cortex. *Proceedings of the National Academy of Sciences*, 102(46), 16875–16879.

Breitmeyer, B. (2007). Visual masking: Past accomplishments, present status, future developments. *Advances in Cognitive Psychology/University of Finance and Management in Warsaw*. http://versita.metapress.com/index/51171305G8L44517.pdf.

Breitmeyer, B., Hoar, W. S., Randall, D. J., & Conte, F. P. (1984). *"Visual masking: An integrative approach*. Clarendon Press.

Brogaard, B. (2015). Type 2 blindsight and the nature of visual experience. *Consciousness and Cognition*, 32, 92–103. Elsevier Inc.

Charles, L., Gaillard, R., Amado, I., Krebs, M.-O., Bendjemaa, N., & Dehaene, S. (2016). Conscious and unconscious performance monitoring: Evidence from patients with schizophrenia. *NeuroImage*, (September)http://dx.doi.org/10.1016/j.neuroimage.2016.09.056.

Charles, L., King, J.-R., & Dehaene, S. (2014). Decoding the dynamics of action, intention, and error detection for conscious and subliminal stimuli. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, 34(4), 1158–1170.

Cowey, A., & Stoerig, P. (1991). The neurobiology of blindsight. *Trends in Neurosciences*, 14(4), 140–145.

Cowey, A., & Stoerig, P. (1995). Blindsight in monkeys. *Nature*, 373, 247–249.

Cowey, A., & Stoerig, P. (1997). Visual detection in monkeys with blindsight. *Neuropsychologia*, 35(7), 929–939.

Deneve, S. (2012). Making decisions with unknown sensory reliability. *Frontiers in Neuroscience*, 6(June), 75.

Eriksen, C. W. (1960). Discrimination and learning without awareness: A methodological survey and evaluation. *Psychological Review*. http://dx.doi.org/10.1037/h0041622. US: American Psychological Association.

Evans, S., & Azzopardi, P. (2007). Evaluation of a 'Bias-Free' measure of awareness. *Spatial Vision*, 20(1), 61–77.

Fetsch, C. R., Kiani, R., Newsome, W. T., & Shadlen, M. N. (2014). Effects of cortical microstimulation on confidence in a perceptual decision. *Neuron*, 1–8. August. Elsevier Inc.

Fleming, S. M., & Daw, N. D. (2017). Self-evaluation of decision-making: A general Bayesian framework for metacognitive computation. *Psychological Review*, 124(1), 91–114. psycnet.apa.org.

Fleming, S. M., & Dolan, R. J. (2012). The neural basis of metacognitive ability. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 367(1594), 1338–1349.

Fleming, S. M., Dolan, R. J., & Frith, C. D. (2012). Metacognition: Computation, biology and function. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 367(1594), 1280–1286.

Fleming, S. M., Huijgen, J., & Dolan, R. J. (2012). Prefrontal contributions to metacognition in perceptual decision making. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, 32(18), 6117–6125.

Fleming, S. M., Maniscalco, B., Ko, Y., Amendi, N., Ro, T., & Lau, H. (2015). Action-specific disruption of perceptual confidence. *Psychological Science*, 26(1), 89–98.

Fogelson, S. V., Kohler, P. J., Miller, K. J., Granger, R., & Tse, P. U. (2014). Unconscious neural processing differs with method used to render stimuli invisible. *Frontiers in Psychology*, 5(June), 601.

Foley, R. (2015). The case for characterising Type-2 blindsight as a genuinely visual phenomenon. *Consciousness and Cognition*, 32, 56–67. Elsevier Inc.

Foley, R., & Kentridge, R. W. (2015). Type-2 blindsight: Empirical and philosophical perspectives. *Consciousness and Cognition*, 32, 1–5.

Gardelle, V. de, & Mamassian, P. (2014). Does confidence use a common currency across two visual tasks? *Psychological Science*, 25(6), 1286–1288.

Giles, N., Lau, H., & Odegaard, B. (2016). What type of awareness does binocular rivalry assess? *Trends in Cognitive Sciences*, 20(10), 719–720.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: John Wiley & Sons, Inc.

Hannula, D. E., Simons, D. J., & Cohen, N. J. (2005). Imaging implicit perception: Promise and pitfalls. *Nature Reviews. Neuroscience*, 6(3), 247–255.

Hedges, J. H., Stocker, A. A., & Simoncelli, E. P. (2011). Optimal inference explains the perceptual coherence of visual motion stimuli. *Journal of Vision*, 11(6), 1–16.

Kammer, T., Puls, K., Strasburger, H., Jeremy Hill, N., & Wichmann, F. A. (2005). Transcranial magnetic stimulation in the visual system. I. The psychophysics of visual suppression. *Experimental Brain Research. Experimentelle Hirnforschung. Experimentation Cerebrale, 160*(1), 118–128.

Kentridge, R. W. (2015). What is it like to have type-2 Blindsight? Drawing inferences from residual function in type-1 blindsight. *Consciousness and Cognition, 32*, 41–44. Elsevier Inc.

Kentridge, R. W., Heywood, C. A., & Weiskrantz, L. (1999). Attention without awareness in blindsight. *Proceedings of the Royal Society B: Biological Sciences, 266*(July), 1805–1811.

Kentridge, R. W., Heywood, C. A., & Weiskrantz, L. (2004). Spatial Attention Speeds Discrimination without Awareness in Blindsight. *Neuropsychologia, 42*, 831–835.

King, J.-R., & Dehaene, S. (2014). A model of subjective report and objective discrimination as categorical decisions in a vast representational space. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 369*(20130204).

Knill, D. C., & Pouget, A. (2004). The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends in Neurosciences*. http://linkinghub.elsevier.com/retrieve/pii/S0166223604003352.

Knill, D. C., & Richards, W. (1996). *Perception as Bayesian inference*. Cambridge University Press.

Ko, Y., & Lau, H. (2012). A detection theoretic explanation of blindsight suggests a link between conscious perception and metacognition. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 367*(1594), 1401–1411.

Kolb, F. C., & Braun, J. (1995). Blindsight in normal observers. *Nature, 377*, 336–338.

Kunimoto, C., Miller, J., & Pashler, H. (2001). Confidence and accuracy of near-threshold discrimination responses. *Consciousness and Cognition, 10*, 294–340. Elsevier Inc.

Kwon, Oh-S., & Knill, D. C. (2013). The brain uses adaptive internal models of scene statistics for sensorimotor estimation and planning. *Proceedings of the National Academy of Sciences, 110*(11), E1064–E1073.

Lau, H. (2007). A higher order bayesian decision theory of consciousness. *Progress in Brain Research, 168*, 35–48.

Lau, H., & Passingham, R. E. (2006). Relative blindsight in normal observers and the neural correlate of visual consciousness. *Proceedings of the National Academy of Sciences, 103*(49), 18763–18768.

Lau, H., & Rosenthal, D. (2011). Empirical support for higher-order theories of conscious awareness. *Trends in Cognitive Sciences, 15*(8), 365–373.

Lloyd, D. A., Abrahamyan, A., & Harris, J. A. (2013). Brain-stimulation induced Blindsight: Unconscious vision or response bias? *PLoS One, 8*(12), 1–16.

Macmillan, N. A., & Douglas Creelman, C. (2004). *Detection theory: A user's guide*. Taylor & Francis.

Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition, 21*(1), 422–430. Elsevier Inc.

Maniscalco, B., & Lau, H. (2016). The signal processing architecture underlying subjective reports of sensory awareness. *Neuroscience of Consciousness*, (November 2015), 1–41.

Maniscalco, B., Peters, M. A. K., & Lau, H. (2016). Heuristic use of perceptual evidence leads to dissociation between performance and metacognitive sensitivity. *Attention, Perception & Psychophysics*. http://dx.doi.org/10.3758/s13414-016-1059-x.

Merikle, P. M., Smilek, D., & Eastwood, J. D. (2001). Perception without awareness : Perspectives from cognitive psychology. *Cognition, 79*, 115–134.

Overgaard, M., & Sandberg, K. (2012). Kinds of access: Different methods for report reveal different kinds of metacognitive access. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 367*(1594), 1287–1296.

Peters, M. A. K., & Lau, H. (2015). Human observers have optimal introspective access to perceptual processes even for visually masked stimuli. *eLife, 10*, 7554/eLife.09651.

Peters, M. A. K., Ro, T., & Lau, H. (2016). Who's afraid of response bias? *Neuroscience of Consciousness, 2016*(1), niw001.

Qamar, A. T., James Cotton, R., George, R. G., Beck, J. M., Prezhdo, E., Laudano, Allison, et al. (2013). Trial-to-Trial, uncertainty-based adjustment of decision boundaries in visual categorization. *Proceedings of the National Academy of Sciences, 110*(50), 20332–20337.

Rahnev, D., Bahdo, L., de Lange, F. P., & Lau, H. (2012a). Prestimulus hemodynamic activity in dorsal attention network is negatively associated with decision confidence in visual perception. *Journal of Neurophysiology, 108*(5), 1529–1536.

Rahnev, D., Maniscalco, B., Graves, T., Huang, E., de Lange, F. P., & Lau, H. (2011). Attention induces conservative subjective biases in visual perception. *Nature Neuroscience, 14*(12), 1513–1515. Nature Publishing Group.

Rahnev, D., Maniscalco, B., Luber, B., Lau, H., & Lisanby, S. H. (2012b). Direct injection of noise to the visual cortex decreases accuracy but increases decision confidence. *Journal of Neurophysiology, 107*, 1556–1563.

Ramsøy, T. Z., & Overgaard, M. (2004). Introspection and subliminal perception. *Phenomenology and the Cognitive Sciences, 3*(1), 1–23.

Ro, T. (2008). Unconscious vision in action. *Neuropsychologia, 46*(1), 379–383.

Ro, T., Breitmeyer, B., Burton, P., Singhal, N., & Lane, D. (2003). Feedback contributions to visual awareness in human occipital cortex. *Current Biology, 11*, 1038–1041.

Ro, T., Dominique, S., Lee, O. L., & Chang, E. (2004). Extrageniculate mediation of unconscious vision in transcranial magnetic stimulation-induced blindsight. *Proceedings of the National Academy of Sciences, 101*(26), 9933–9935.

Rounis, E., Maniscalco, B., Rothwell, J. C., Passingham, R. E., & Lau, H. (2010). Theta-burst transcranial magnetic stimulation to the prefrontal cortex impairs metacognitive visual awareness. *Cognitive Neuroscience, 1*(3), 165–175.

Sahraie, A., Hibbard, P. B., Trevethan, C. T., Ritchie, K. L., & Weiskrantz, L. (2010). Consciousness of the first order in blindsight. *Proceedings of the National Academy of Sciences, 107*(49), 21217–21222.

Sandberg, K., Timmermans, B., Overgaard, M., & Cleeremans, A. (2010). Measuring consciousness: Is one measure better than the other. *Consciousness and Cognition*. Cognitive Neuroscience Research Unit, Hammel Neurorehabilitation and Research Center, Denmark: Elsevier Inc. http://linkinghub.elsevier.com/retrieve/pii/S1053-8100(09) 00199–8

Seriès, P., Stocker, A. A., & Simoncelli, E. P. (2009). Is the homunculus 'aware' of sensory adaptation. *Neural Computation, 21*, 3271–3304.

Stocker, A. A., & Simoncelli, E. P. (2006). Noise characteristics and prior expectations in human visual speed perception. *Nature Neuroscience, 9*(4), 578–585.

Stocker, A. A., & Simoncelli, E. P. (2008). A Bayesian model of conditioned perception. *Advances in Neural Information Processing Systems, 20*(May), 1409–1416.

Vilares, I., Howard, J. D., Fernandes, H. L., Gottfried, J. A., & Körding, K. P. (2012). Differential representations of prior and likelihood uncertainty in the human brain. *Current Biology: CB, 22*(18), 1641–1648.

Vilares, I., & Körding, K. P. (2011). Bayesian models: The structure of the world, uncertainty, behavior, and the brain. *Annals of the New York Academy of Sciences, 1224*(April), 22–39.
Weiskrantz, L. (1986). *Blindsight: A case study and implications*. Oxford University Press.
Weiskrantz, L. (1996). Blindsight revisited. *Current Opinion in Neurobiology, 6*, 215–220.
Yuille, A. L., & Bülthoff, H. H. (1996). *Bayesian decision theory and psychophysics*. In D. C. Knill, & W. Richards (Eds.) (pp. 123–161). New York: Cambridge University Press.
Zylberberg, A., Fetsch, C. R., Shadlen, M. N., & Frank, M. J. (2016). The influence of evidence volatility on choice, reaction time and confidence in a perceptual decision. *eLife, 5*(October), e17688. eLife Sciences Publications Limited.
Zylberberg, A., Roelfsema, P. R., & Sigman, M. (2014). Variance misperception explains illusions of confidence in simple perceptual decisions. *Consciousness and Cognition, 27C*(June), 246–253. Elsevier Inc.