Counterfactual Reasoning in Perceptual Decision-Making

# Introduction

Metacognitive sensitivity refers to a person's ability to judge their own accuracy in psychophysical tasks through their degree of confidence. This metric can be quantified as the extent to which confidence can discriminate between incorrect and correct trials (Fleming & Lau, 2014). In perceptual detection tasks, metacognitive sensitivity for judgments about absence (‘no’ responses) is lower than for judgments about presence (‘yes’ responses; Meuwese, van Loon, Lamme, and Fahrenfort, 2014; Kanai, Walsh, and Tseng, 2010). This disadvantage for ‘no’ responses is partly recovered when task difficulty is controlled using attentional, rather than perceptual, experimental manipulations (Kanai, Walsh and Tseng, 2010; Kellij et al., 2018). One possibility is that this is a result of counterfactual reasoning, whereby participants use their knowledge about their current attentional state to reason about the likelihood of detecting a hypothetical target (i.e., “Given my current attention state, I might have missed the target”).

This study aims to generalise this finding to situations where participants have access to external cues that can provide information about the likelihood of presented targets to be detected, based on their expected objective visibility, rather than internal cues, such as attentional state. The term *expected visibility*, which will appear throughout this document, refers to a participant’s expectation about the visibility of a stimulus just before it is presented, based on previous trials. Our main hypothesis for this study is that during target-absent trials, this expectation may influence subjects’ decisions (‘yes’ or ‘no’) and/or subjective confidence. More specifically, we predict that in ‘no’ responses, this expectation will influence confidence ratings via counterfactual reasoning, whereby upon not seeing a target, participants will use beliefs about hypothetical targets to inform their subjective confidence ratings (“Given the expected stimulus visibility, I might have missed the target”).

## Aims and Objectives

This study aims to investigate the role of counterfactual evidence in the decision-making process involved in perceptual detection. Namely, we hope to:

1. Determine whether expected visibility, as estimated based on previous trials, has an effect on response bias in detection.
2. Determine whether expected visibility, as estimated based on previous trials, has an effect on confidence ratings in ‘yes’ responses.
3. Determine whether expected visibility, as estimated based on previous trials, has an effect on confidence ratings in ‘no’ responses. This would test whether participants are able to use expectations about an unobserved stimulus to inform their confidence about not seeing a stimulus.
4. Test the effects of *expected volatility* of visibility fluctuations on these patterns. This will be done by contrasting experimental blocks where stimulus visibility is autocorrelated across trials with blocks in which stimulus visibility changes randomly without any dependence on previous trials.

# Materials and Methods

## Participants

35 healthy participants will take part in the experiment.

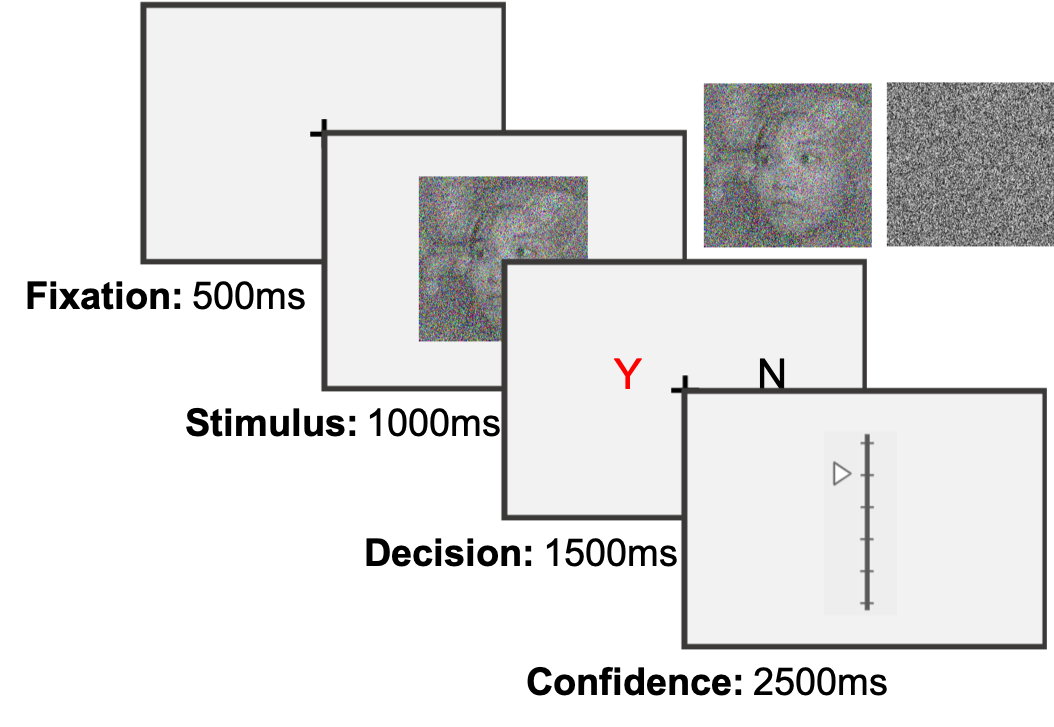
### Experimental Procedure

Each participant will complete a short practice round and then perform four blocks of 120 trials each; two structured blocks followed by two random blocks.

### Trial structure

Trial onset will be indicated by the appearance of a fixation cross (Figure 1). After 500 milliseconds, the stimulus will appear behind the fixation cross for 1000ms. In signal-absent trials, the stimulus will consist of a dynamic array of grayscale pixels, which will change their brightness values psuedorandomly every frame, creating a noise display. In signal-present trials (50% of all trials), a face will emerge from the noise for a brief moment during this one second interval. The noisy pixels will gradually take the structure of the image until it emerges from the noise briefly and then fades shortly thereafter. The peak of the face image will occur anytime between 267 milliseconds and 767 milliseconds into the trial. Face visibility will be controlled by manipulating the proportion of pixels that take non-random brightness values from the face image. For example, in high-visibility trials, the peak-visibility frame could include 28% of ‘face pixels’, and in low-visibility trials the peak proportion of ‘face pixels’ might only be 12%. The set of stimuli used for this experiment consisted of 1,579 images of faces taken from the Flickr-Faces-HQ (FFHQ) Dataset (Karras, Laine and Aila, 2019), transformed into grayscale and compressed to 300 by 300 pixels.

Following the stimulus, a decision screen will be presented with the options “Y” for ‘yes’ and “N” for ‘no’. Participants will have 1.5 seconds to indicate, using their left hand, whether a face was presented or not. The “a” key will indicate YES and the “s” key will indicate NO. Their response will highlight the corresponding letter until the 1.5 second interval terminates. After this, the participants will have 2.5 seconds to rate their confidence in their decision. This will be done using a vertical scale from 1 (lowest confidence) to 6 (highest confidence), with the marker starting at a random location every trial. They will use their right hand to press the “up arrow” key to move the marker up (higher confidence) and the “down arrow” key to move it down (lower confidence).



**Figure 1:** **Schematic representation of the experimental procedure.**

To incentivize participants to perform their best at the task and rate their confidence accurately, a bonus payment will be offered such that:

where is a vector of 1 and -1 for correct and incorrect responses, and is a vector of integers in the range of 1 to 6, representing confidence reports for all trials. The payment structure will be explained to participants. Specifically, participants will be advised that in order to maximize their bonus they should perform their best at the main task, rate their confidence higher when they believe that they were correct, and rate their confidence lower when they believe that they might have been incorrect. Participants will be informed about their bonus at the end of each block.

### Block Structure

The first two blocks of the experiments will be ‘structured blocks’. These blocks will start with a monotonic decrease in stimulus visibility for the first 20 trials, followed by a random walk of stimulus visibility. The random walk will take place in the log-visibility space, since sensitivity to changes is expected to be proportional to the stimulus strength. Stimulus visibility for trials in the structured blocks will thus be autocorrelated over time.

The last two blocks will be ‘random blocks’. Stimulus visibility for these blocks will be determined by shuffling the order of visibility values from the structured blocks. Stimulus visibility for trials in the random blocks will thus not be autocorrelated over time.

### Analysis

Expected visibility will be extracted by taking the stimulus visibility at the last trial in which the participant indicated seeing a stimulus (i.e. a ‘yes’ response). For example, after a trial in which the participant correctly detected the stimulus, which was presented at a visibility of -2, expected visibility for the subsequent trial will be set to -2. This value will only change when the participant responds ‘yes’. In false alarm trials, where the participant incorrectly responds ‘yes’ to an absent stimulus, expected visibility will be set to -5 (the lowest value).

We will test the Spearman correlations between expected visibility and response, and between expected visibility and confidence. We are interested in comparing the correlation with response between structured and random blocks, and the correlation with confidence between responses (yes or no) and blocks (structured or random).

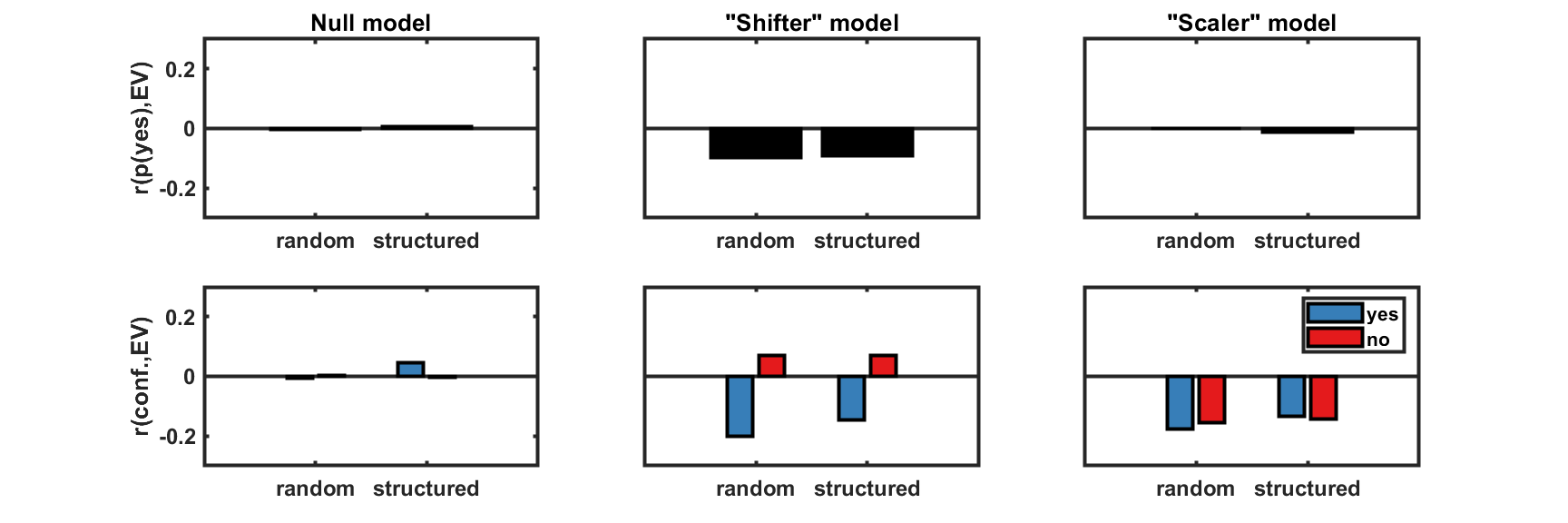
Three theoretical models, all based on Signal Detection Theory (SDT; Wickens 2002), are hereby considered and will be used to derive predictions. These models assume that individual decisions and confidence ratings are based on the position of perceptual samples on a “strength-of-evidence” dimension, such that the mean strength-of-evidence of the signal distribution is greater than that of the noise distribution. The variance of these distributions is composed of external variance in stimulus visibility (e.g. contrast), and internal/perceptual noise. While the second is equal for both distributions, the first is positive only for the signal distribution, and is set to 0 for the noise distribution. In order to make a decision about whether signal or noise was presented on a given trial, the strength-of-evidence of the perceptual sample is compared with a decision criterion. If lower, the subject responds ‘no’, whereas any strength equal to or above the criterion results in a ‘yes’ response. Confidence in a decision increases as a function of the absolute distance of the strength-of-evidence from the decision criterion. This is commonly modeled by including a set of equidistant confidence criteria around the decision criterion.

The first model in this study is considered a “null” model, as it is based on a classic SDT paradigm without any assumptions about the effects of previous counterfactual evidence or expected visibility on subjects’ decisions. The decision criterion of this model is static throughout the experiment and occupies the midpoint between the peaks of the signal and noise distribution, which is simply the average of the signal and noise means. Similarly, the distance between confidence criteria is maintained constant throughout the experiment. Since the model does not change these criteria, it makes no predictions about the effects of expected visibility on either confidence ratings (Figure 2; lower panel) or responses (upper panel). As a consequence of the fact that expected visibility will be correlated with true stimulus visibility in structured blocks, spurious correlations may exist between expected visibility and either confidence or response in these blocks. Importantly, the null model predicts no correlations in random blocks.

The other two models are influenced by expected visibility, which is assumed to reflect the participant’s belief about the mean of the distribution of visibility levels. The first of these is a “shifter” model, which assumes a rigid shift in criteria that is proportional to the expected visibility. In this model, the decision criterion is set to the middle point between the peak of the noise distribution and the expected visibility, leading to more conservative decisions when subjects expect more visible stimuli. The “shifter” model predicts a negative correlation of confidence with expected visibility in “yes” responses, and a positive correlation of confidence with expected visibility in “no” responses. This model implements a simple tracking of external cues (stimulus visibilities on previous trials) to compute counterfactual stimulus strength.

The last model is a “scaler” model, which assumes a scaling in the confidence criteria proportional to the expected visibility, with no change to the decision criterion. In this model, high expected visibility will induce a symmetric spreading of the confidence criteria around the decision criterion, reducing general confidence levels without an effect on decision. This model corresponds to an agent that assumes that signal and noise are symmetric.

The observed patterns will be qualitatively compared with the patterns predicted by our three theoretical models. In particular, we note that only the ‘shifter’ model predicts a negative effect of expected visibility on the probability of responding ‘yes’. Furthermore, while both the ‘shifter’ and ‘scalar’ models predict negative effects of expected visibility on confidence in ‘yes’ responses, only the ‘shifter’ model predicts a positive effect of expected visibility on confidence in ‘no’ responses (Figure 2).

**Figure 2:** **Model predictions for the correlation between expected visibility and decision (upper panels), and for the correlation between expected visibility and confidence (lower panels).**

### Transparency

All code and materials for this project are publicly available on github.com/matanmazor/counterfactualStimuli. This document is uploaded to the open science framework after completing data collection from the first 12 participants.

## References

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in human neuroscience*, *8*, 443.

Kanai, R., Walsh, V., & Tseng, C. H. (2010). Subjective discriminability of invisibility: a framework for distinguishing perceptual and attentional failures of awareness. *Consciousness and cognition*, *19*(4), 1045-1057.

Karras, T., Laine, S., & Aila, T. (2018). A style-based generator architecture for generative adversarial networks. *arXiv preprint arXiv:1812.04948*.

Kellij, S., Fahrenfort, J. J., Lau, H., Peters, M. A. K., & Odegaard, B. (2018, December 13). The foundations of introspective access: how the relative precision of target encoding influences metacognitive performance. <https://doi.org/10.31234/osf.io/xky38>

Meuwese, J. D., van Loon, A. M., Lamme, V. A., & Fahrenfort, J. J. (2014). The subjective experience of object recognition: Comparing metacognition for object detection and object categorization. *Attention, Perception, & Psychophysics*, *76*(4), 1057-1068.

Wickens, T. D. (2002). *Elementary signal detection theory*. Oxford University Press, USA