An Ideal Observer Model Of Audiovisual Detection Captures Modality-Specific, But Not Amodal, Confidence Ratings

Perrine Porte (perrine.porte@univ-grenoble-alpes.fr)
Université Grenoble Alpes, CNRS, LPNC
Matan Mazor (matan.mazor@all-souls.ox.ac.uk)
University of Oxford
Louise Goupil (louise.goupil@univ-grenoble-alpes.fr)
Université Grenoble Alpes, CNRS, LPNC
Nathan Faivre (nathan.faivre@univ-grenoble-alpes.fr)
Université Grenoble Alpes, CNRS, LPNC

Abstract

Detecting objects in the environment, and forming a sense of confidence in these decisions typically involves multisensory processing. We sought to characterize how humans form amodal and modality-specific confidence judgments during audiovisual detection. We found that participants made more accurate detection and confidence judgments for audiovisual than unimodal stimuli. To explain these results, we extended a Bayesian evidence accumulation model to audiovisual detection and successfully reproduced both unimodal and audiovisual detection judgments. Despite being fitted to decisions and decision times alone, our model accurately reproduced modality-specific confidence. It failed, however, to account for amodal confidence, suggesting that the latter might not arise from optimal signal integration in detection contexts. Our results indicate that, in the presence of audiovisual signals, different integration rules apply for perceptual and metacognitive decisions.

Keywords: metacognition; multisensory integration; detection decision; Bayesian modelling; audiovisual perception.

Introduction

Imagine trying to decide if a mosquito is present before going to sleep. If you hear it, are you more confident that there is one if you also see it? And if you neither hear it nor see it, how confident are you that there is no mosquito?

The impact of multisensory signals has been vastly studied at the perceptual level — showing that people have better performance and faster reaction times in the presence of bimodal stimuli compared to unimodal ones (Rach et al., 2011; Plass & Brang, 2021) — but it remains poorly described at the meta-perceptual level (Deroy et al., 2016). Recent studies showed that, despite an improvement in task performance, metacognitive performance is not better in multimodal compared to unimodal conditions (Charles et al., 2020; Arbuzova et al., 2021; Faivre et al., 2018). However, these studies only used discrimination tasks, and metacognitive processes may differ between discrimination and detection tasks (Mazor et al., 2020; Mazor et al., 2023).

Evidence accumulation models have been developed to explain multisensory effects on detection performance. They assume that sensory evidence is accumulated from each modality until a decision threshold is reached, with decision times reflecting the time to reach the threshold. Accumulation rates from different modalities are typically considered additive, and correlated (Plass & Brang, 2021; Diederich, 1995; Gondan et al., 2010). These models typically account for faster decision times in bimodal than in unimodal conditions, but provide limited insight into decision accuracy. In addition, they do not consider people's ability to monitor the modality through which the stimulus was detected. To address this, Blurton et al. (2014) introduced a drift-diffusion model with two thresholds, allowing predictions about both decision times and accuracy. However, this model was based on an identification task, where participants needed to detect a specific stimulus and ignore distractors. Since a stimulus was always present, no comparisons could be made between decisions about the presence versus absence of sensory evidence: judgments that are known to rely on partly different mechanisms.

Indeed, in a recent study using visual stimuli, Mazor et al. (2024) showed that optimal decision-making in detection tasks relies on the integration of factual and counterfactual evidence, where participants only accumulated evidence for presence (i.e., factual evidence) and inferred absence from counterfactual detectability (i.e., "*I would have perceived it if it was present*"). This approach differs from previous

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

evidence accumulation models which typically assume that the same accumulation process similarly drives decisions about presence and absence. In this work, we aim to extend such a model to explain what information is integrated into confidence when detecting audiovisual stimuli.

Method

Sixty participants performed a pre-registered experiment (https://osf.io/3nvyx) audiovisual detection task. Participants observed dynamic visual Gaussian noise while listening to auditory pink noise. A visual stimulus (a light gray circle) was embedded in visual noise on half of the trials. Independently, an auditory stimulus (a sinusoidal tone of 1 kHz) was embedded in auditory noise on half of the trials. As a result, a signal was present on 75% of the trials, with 25% of the trials including both a visual and an auditory signal presented simultaneously (AV trials), and 25% of the trials including none (absent trials; see Fig. 1). At the beginning of the experiment, stimulus intensity was adjusted for each participant to reach a 40% detection rate in visual and auditory conditions, based on unimodal psychometric curves. Participants then undertook the main task. The stimulus, if present, could appear 200, 300, or 400 ms after the noise onset and was presented for 600 ms. Participants could respond right from the start of the trial until a limit of 4s after the stimulus offset. On each trial, they indicated whether a stimulus was present or absent irrespective of its sensory modality, before reporting their amodal confidence regarding their detection choice from 0 ("sure incorrect") to 100 ("sure correct"). Critically, they were instructed to report presence if a stimulus was present visually, auditorily, or both. Likewise, they were asked to report an amodal confidence judgement reflecting decision accuracy irrespective of the sensory modality. Finally, they reported their modality-specific detection and confidence judgments on a bi-dimensional (audio/visual) report scale, with each axis ranging from 100% sure not perceived to 100% sure perceived, and corresponding to one modality (Fig. 1). Participants performed a total of 288 trials (72 per each of the four conditions: A, V, AV and absent).

Based on our pre-registered exclusion criteria (no variability in confidence judgments or no convergence of psychometric curves), three participants were excluded. Additionally, six participants were excluded as their hit rate was 0% in either the auditory or the visual modality. As trials with confidence ratings below 50 were not analyzed, three participants were excluded as they responded with a confidence below 50 in more than half of the trials. 48 participants were included in the main analysis.

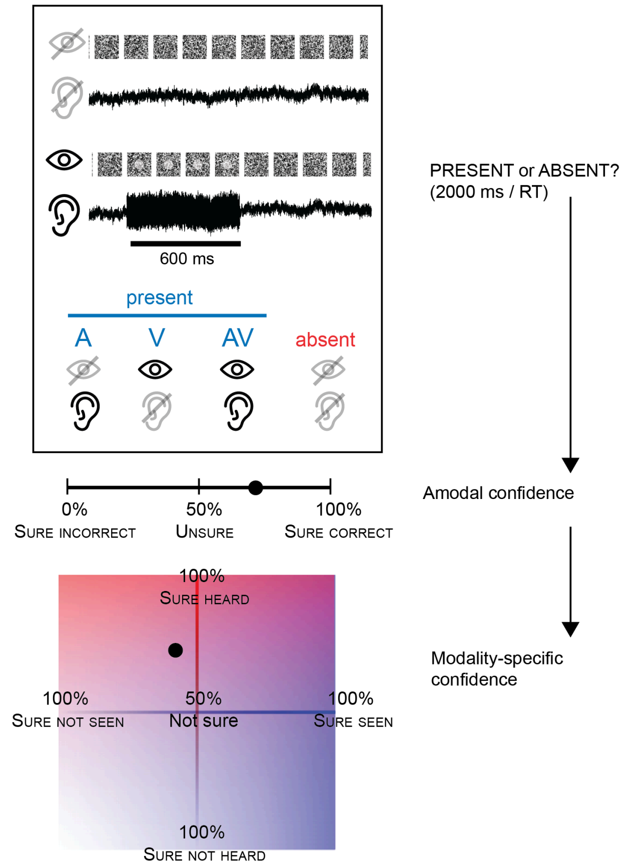

Figure 1: Trial structure. In frame: different types of stimuli that could be presented.

Model

We extended a recent ideal-observer model of visual detection to account for our multimodal detection task (Mazor et al., 2024). In the original model, agents observe the activation of a single "presence sensor". At each time point, this sensor is either sampling a 1 or a 0, and the probability of sampling a 1 is higher if the target is present. Sensor activation probabilities are captured by model parameters $\theta_{present}$ (i.e., probability of sampling a 1 if a target is present) and θ_{absent} (i.e., probability of sampling a 1 if a target is absent). Importantly, this model assumes that agents hold beliefs about these probabilities (i.e., beliefs about the probability of sampling a 1, if the target is present or absent). These beliefs are captured by model parameters $\bar{\theta}_{present}$ and $\bar{\theta}_{absent}$, and reflect the degree to which agents believe that they would have perceived the target if it was present. The output decision is based on the log-likelihood ratio of the two competing hypotheses (present or absent, Mazor et al., 2024).

To adapt this model for audiovisual detection, we included two modality-specific sensors: a visual sensor and an auditory sensor (Fig. 2). We adopted a disjunctive rule as a rational integration rule, such that *p(x or y)=p(x)+p(y)-p(x and y)*, reflecting that a stimulus can be (objectively) present in one modality or in both. The model also allowed different prior beliefs regarding the probability of presence in the visual and auditory modalities. While being fitted to amodal detection and decision time data alone, our model made qualitative predictions about amodal confidence based on the probability of being correct at the time of the decision. Furthermore, it made predictions about modality-specific effects as it computed the probability of presence separately for each modality: a stimulus was judged present if the modality-specific probability of presence was higher than 0.5 at the time of the decision, and absent otherwise. Modality-specific confidence was also read as the probability of being correct at the time of the decision, for each modality separately.

We compared the goodness of fit to amodal decisions for different combinations of parameters representing the prior and likelihood (Table 1). Model comparisons showed that the best model was the “single prior” model for which all parameters, except the prior, varied across modalities.

Table 1: Fitted models (rows) with different or identical parameters (columns) across the visual and auditory sensors.

	Prior	True thetas θ	Believed thetas $\bar{\theta}$	AIC (relative to the single prior model)
Single prior	Same	Different	Different	0
Full	Different	Different	Different	109
Single belief	Different	Different	Same	132
No belief	Same	Different	None	318
Single likelihood and single belief	Different	Same	Same	1111
No belief and single likelihood	Same	Same	None	1404

To assess the model’s ability to reproduce qualitative patterns in the data, we simulated data from 48 agents based on parameters fitted to individual participants, using the single prior model.

Statistical Analysis

To investigate reaction times (in seconds), we performed a mixed-effect linear regression on the log of the reaction

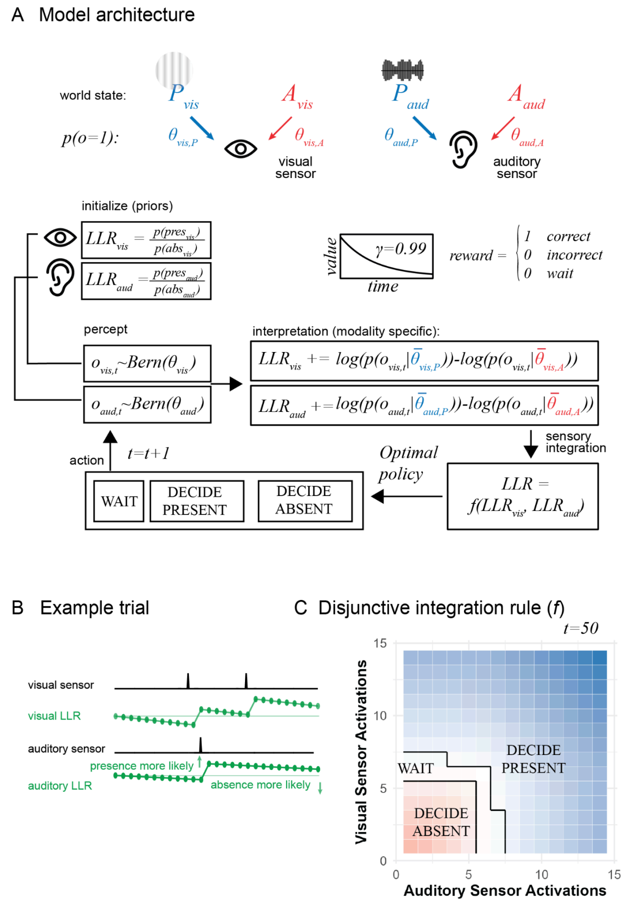

Figure 2: Computational model. A) Model architecture. The observer is assumed to have access to a visual sensor and an auditory sensor, probabilistically tuned to the presence of visual and auditory evidence. The probability of activation is controlled by the theta parameter. The agent observes the activations and updates its beliefs about the presence of a signal in each modality separately, using Bayes’ rule. The agent then integrates the two beliefs into an amodal belief in the presence of a target. Based on this belief, it decides whether to commit to a decision or accumulate more evidence by following an optimal policy, derived using backward induction (Callaway et al., 2023). B) Example trial: modality-specific LLRs (in green) are updated following sensor inactivations and activations. C) The disjunctive integration rule. Amodal LLR is plotted as a function of the number of sensor activations in each modality 50 time points into the trial. Black contours indicate regions in which the best action is to decide present, wait, or decide absent.

time, with the modality, the accuracy of the responses (correct or incorrect), and their interaction both as fixed and random effects. To investigate detection accuracy and metacognitive sensitivity we performed a mixed-effect logistic regression on the participants' accuracy (correct or incorrect) with the modality of presentation, confidence and their interaction both as fixed effect and random effect, for both amodal and modality-specific judgments.

Results

Better Detection Of Audiovisual Stimuli

Participants had a mean d' of 1.5 (*SD* = 0.64) and a significant bias for responding "absent" with a positive SDT criterion of 0.58 (*SD* = 0.43, $t(47) = 9.32$, $p < .001$). Participants were more accurate for bimodal (*mean proportion correct* = 0.74, *SD* = 0.44) compared to unimodal trials (*mean proportion correct* = 0.51, *SD* = 0.5) ($p < .001$), and for visual (*mean proportion correct* = 0.6, *SD* = 0.49) compared to auditory trials (*mean proportion correct* = 0.42, *SD* = 0.49; $p < .001$; Fig. 3A). This pattern is consistent across measures (i.e., same results with d' comparison). The model successfully reproduced amodal detection behavior, with a mean d' of 1.51 (*SD* = 0.66) and a bias for responding "absent" with a positive SDT criterion of 0.57 (*SD* = 0.42). It also reproduced the higher accuracy for audiovisual compared to unimodal trials, and for visual compared to auditory trials (Fig. 3A).

Participants were faster when responding accurately (*M* = 1.45s, *SD* = 0.43) compared to inaccurately (*M* = 1.5s, *SD* = 0.51, $p < .001$). A significant interaction between detection accuracy, stimulus presence, and stimulus modality showed that the effect of accuracy on decision time was stronger in present compared to absent trials ($p = .004$), for bimodal compared to unimodal trials ($p = .002$), and for visual compared to auditory trials ($p = .002$) (Fig. 3B). Our model reproduced this result, but for correct answers only (Fig. 3B).

Participants Correctly Identified Stimulus Modality

Critically, the bi-dimensional report scale allowed us to ask how accurate participants were in their modality-specific detection of visual and auditory stimuli. For correct target-presence judgments only, we examined whether participants accurately categorized the modality in which the stimulus was presented—a form of source monitoring ("*did I see or hear this stimulus?*"). Participants accurately categorized the source of their percept, with respectively 88% and 75% of auditory and visual stimuli being

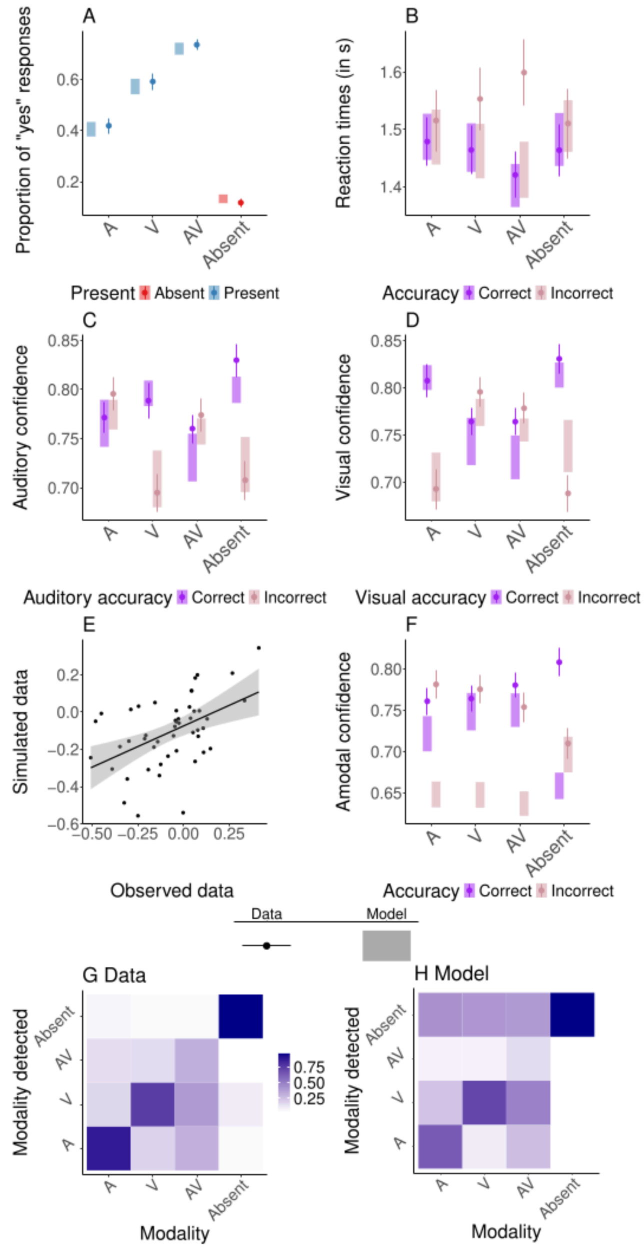

Figure 3: Results. Error bars represent the standard error from the data. Rectangles represent data simulated from the model, centered on the mean value and with height equal to the standard error. A) Percentage of "yes" responses B) Reaction times (in seconds) for the amodal detection. C) Auditory confidence. D) Visual confidence. E) Correlation between observed and simulated data for the confidence asymmetry index. F) Amodal confidence. G-H) Modality detected by modality for correct answers only.

categorized as such. For audiovisual stimuli, participants detected them as being auditory (*M* = 31%, *SD* = 4%), visual (*M* = 38%, *SD* = 4%) or audiovisual (*M* = 31%, *SD* =

3%). These lower figures are expected if detecting stimuli as audiovisual is based on the joint probability of detecting them visually (hit rate in V trials: 60%) *and* auditorily (hit rate in A trials: 42%). Interestingly, our model predicted that 42% of trials accurately judged as being present are judged to be absent in both modalities (Fig. 3H), compared with only 3% in the data from actual participants (Fig. 3G), meaning that the model, but not the participants, could judge that something was present without judging that the stimulus was present in at least one of the two modalities.

An Advantage For Audiovisual Stimuli In Modality-Specific Metacognition

Turning to modality-specific confidence, we defined metacognitive sensitivity as the ability to distinguish between correct and incorrect responses based on confidence judgments (Fleming & Lau, 2014). We focused on modality-specific target-present trials (i.e., AV and A trials when analysing the auditory modality, and AV and V trials when analysing the visual modality). Metacognitive sensitivity was approximated from the slope of the logistic regression predicting modality-specific accuracy (hits and misses) from confidence. We found that modality-specific metacognitive sensitivity was higher for audiovisual compared to unimodal trials for both the auditory (p = .025) and the visual modalities (p = .008) (Fig. 3C-D).

Despite being fitted to amodal decision and decision times only, our model captured the observed modality-specific confidence effects with a simulated modality-specific metacognitive sensitivity higher for audiovisual than unimodal trials for both the auditory and the visual modalities (Fig. 3C-D). Moreover, the model captured between-participant variability in confidence asymmetries between the auditory and visual modalities, for the same (AV and absent) stimuli. Specifically, for each participant, we defined and examined a "confidence asymmetry index": (auditory confidence in AV trials−auditory confidence in absent trials)−(visual confidence in AV trials−visual confidence in absent trials). A significant positive correlation was observed between the observed and simulated confidence asymmetry indices (Pearson's *rho* = .68, p < .001), indicating that the model successfully captured the interindividual variability of visual and auditory weights on confidence (Fig. 3E).

An Advantage For Audiovisual Stimuli In Amodal Metacognition

Next, we turned to amodal confidence ratings, again taking the slope of the logistic regression predicting accuracy (hits and misses) from confidence as an index of metacognitive sensitivity for presence. We found that amodal metacognitive sensitivity was higher for bimodal compared to unimodal trials (p < .001). There was no significant difference between visual and auditory trials (Fig. 3F). This shows that audiovisual stimuli were not only more likely to be detected; they were also followed by a tighter coupling between amodal confidence and objective accuracy (i.e. afforded greeted amodal metacognitive sensitivity).

In contrast to its overall success in predicting modality-specific confidence ratings, the model failed to predict the amodal confidence effects. It produced an overall higher confidence for presence judgments that was not observed behaviorally, and failed to reproduce the higher metacognitive sensitivity for audiovisual stimuli compared to unimodal ones (Fig. 3F).

Predicting Amodal Confidence From Modality-Specific Confidence Ratings

We reasoned that the model's failure might be due to different integration rules applied at the perceptual and metacognitive levels. To investigate whether modality-specific confidence could predict amodal confidence, we combined the raw detection judgments and confidence ratings to one scale, going from full certainty in absence to full certainty in presence (hereafter *detection ratings)*. We did so separately for amodal and modality-specific decisions. We compared different models: 1) a linear model predicting amodal ratings by a weighted linear combination of the visual and auditory rating; 2) a disjunctive model predicting amodal ratings by a linear combination of the visual and auditory ratings minus the product of the two; 3) a min-max model predicting amodal ratings by a weighted linear combination of the highest and lowest modality-specific ratings. This model could express the heuristic of, for example, basing the amodal rating predominantly on the more clearly perceived modality. We found that the best model was the min-max model, indicating that what drives amodal confidence is not a combination of modality-specific confidence ratings but rather, the highest and lowest ratings on a given trial. This fit was further improved by taking into account the absolute confidence of participants separately for presence and absence judgments. To compute this absolute confidence we took the mean of the distance to 0.5 across modalities (i.e., absolute(auditory confidence - 0.5) + absolute(visual confidence - 0.5)). Inspection of model coefficients revealed a main effect of the maximal confidence (b = 0.15, *95%CI* [0.1,0.2], p < .001) and a smaller effect of the minimal confidence ($b = 0.08$, *95%CI* [0.05,0.11], $p < .001$) with an interaction effect between the type of judgments and the absolute confidence ($b = 0.96$, *95%CI* [0.79,1.12], $p < .001$). Applying subject-wise coefficients to simulated data from the ideal observer model reproduced amodal confidence in unimodal trials, but failed to account for the increase in metacognitive sensitivity we observed in AV trials (Fig. 4).

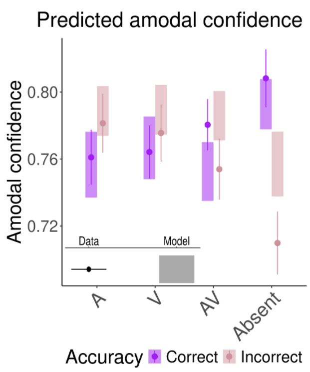

Figure 4: Predicted amodal confidence by the min-max with absolute confidence model. Error bars show standard error, and rectangles represent model-simulated data, centered on the mean values with height equal to the standard error. Based on observed subject-wise coefficients, we predicted the amodal detection ratings from the simulated modality-specific detection ratings.

Discussion

The redundancy of sensory evidence across multiple sensory channels is known to increase task performance (Rach et al., 2011; Plass & Brang, 2021; Blurton et al., 2014). However, how it impacts confidence judgments, especially in a detection task, remains an open question (Deroy et al., 2016). To answer this question, we developed a paradigm in which participants detected audiovisual stimuli before providing both amodal and modality-specific confidence ratings. To model this data, we adapted a recent Bayesian evidence accumulation framework assuming that absence is inferred based on counterfactual detectability (Mazor et al., 2024). By adding a disjunctive rule to integrate information from the different modalities, we reproduced key aspects of audiovisual detection and confidence: our model captured the higher amodal detection performance for bimodal compared to unimodal stimuli, a typical finding in bimodal detection (e.g., Schnupp et al., 2005; Gondan et al., 2004; Plass & Brang, 2021). Moreover, when looking more closely at correct detection answers, our model, similar to participants, could identify the modality in which the stimulus was perceived. A curious discrepancy between human and model behaviour emerged in modality-specific detection judgments following an amodal “target-present” response. In our data, when participants made an amodal judgment that a stimulus was present, they almost always reported it as present in at least one modality. In contrast, the model could reach a “present” decision without any modality having reached a “present” decision independently. This sheds light on an interesting property of human perception: participants, unlike the Bayes-rational model, strived to be consistent in their amodal and modality-specific detection judgments. This consistency could be achieved by making a “present” decision only once presence is sufficiently likely in at least one of the modalities. Alternatively, humans may retrospectively assign a higher probability of presence to one of the two modalities, in an attempt to minimise cognitive dissonance.

Despite not being fitted to confidence ratings, our model was able to reproduce the modality-specific confidence ratings for the two modalities. This result corroborates previous findings showing that, in some settings, confidence closely matches the probability of being correct, especially when decision and confidence are reported simultaneously (Pouget et al., 2016; Aitchison et al., 2015).

Our model, however, did not reproduce participants’ amodal confidence ratings. For example, it systematically predicted a higher confidence for presence than for absence. Additional analysis showed that amodal ratings weigh modality-specific evidence according to their relative strength irrespective of the sensory modality. The absolute confidence participants exhibited in their modality-specific ratings — representing fluctuations in confidence from trial to trial regardless of detection — also contributed to amodal confidence. Future work will be needed to determine which additional features of evidence accumulation could explain amodal confidence, like post-decisional mechanisms (Msheik et al., 2025).

To conclude, an ideal observer model, equipped with two sensors and a disjunctive integration rule, successfully reproduced amodal and modality-specific detection performance, and modality-specific confidence ratings, but not amodal confidence ratings. Taken together with results from additional analysis, this suggests that different integration rules apply to perceptual and metacognitive decisions.

Acknowledgments

This work is supported by the ERC grant (Volta, 101125379) awarded to NF. MM is supported by a postdoctoral fellowship from All Souls College at the University of Oxford. The authors thank the IDEX for funding PP mobility grant.

References

Aitchison, L., Bang, D., Bahrami, B., & Latham, P. E. (2015). Doubly Bayesian Analysis of Confidence in Perceptual Decision-Making. *PLOS Computational Biology*, *11*(10), e1004519. https://doi.org/10.1371/journal.pcbi.1004519

Arbuzova, P., Peters, C., Röd, L., Koß, C., Maurer, H., Maurer, L. K., Müller, H., Verrel, J., & Filevich, E. (2021). Measuring metacognition of direct and indirect parameters of voluntary movement. *Journal of Experimental Psychology: General*, *150*(11), 2208‑2229. https://doi.org/10.1037/xge0000892

Blurton, S. P., Greenlee, M. W., & Gondan, M. (2014). Multisensory processing of redundant information in go/no-go and choice responses. *Attention, Perception, & Psychophysics*, *76*(4), 1212‑1233. https://doi.org/10.3758/s13414-014-0644-0

Callaway, F., Griffiths, T. L., Norman, K. A., & Zhang, Q. (2023). Optimal metacognitive control of memory recall. Psychological Review, No Pagination Specified–No Pagination Specified. http://doi.org/10.1037/rev0000441

Charles, L., Chardin, C., & Haggard, P. (2020). Evidence for metacognitive bias in perception of voluntary action. *Cognition*, *194*, 104041. https://doi.org/10.1016/j.cognition.2019.104041

Deroy, O., Spence, C., & Noppeney, U. (2016). Metacognition in Multisensory Perception. *Trends in Cognitive Sciences*, *20*(10), 736‑747. https://doi.org/10.1016/j.tics.2016.08.006

Diederich, A. (1995). Intersensory Facilitation of Reaction Time : Evaluation of Counter and Diffusion Coactivation Models. *Journal of Mathematical Psychology*, *39*(2), 197‑215. https://doi.org/10.1006/jmps.1995.1020

Faivre, N., Filevich, E., Solovey, G., Kühn, S., & Blanke, O. (2018). Behavioral, Modeling, and Electrophysiological Evidence for Supramodality in Human Metacognition. *Journal of Neuroscience*, *38*(2), 263‑277. https://doi.org/10.1523/JNEUROSCI.0322-17.2017

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience*, *8*, 443. https://doi.org/10.3389/fnhum.2014.00443

Gondan, M., Götze, C., & Greenlee, M. W. (2010). Redundancy gains in simple responses and go/no-go tasks. *Attention, Perception, & Psychophysics*, *72*(6), 1692‑1709. https://doi.org/10.3758/APP.72.6.1692

Gondan, M., Lange, K., Rösler, F., & Röder, B. (2004). The redundant target effect is affected by modality switch costs. *Psychonomic Bulletin & Review*, *11*(2), 307‑313. https://doi.org/10.3758/BF03196575

Mazor, M., Friston, K. J., & Fleming, S. M. (2020). Distinct neural contributions to metacognition for detecting, but not discriminating visual stimuli. *eLife*, *9*, e53900. https://doi.org/10.7554/eLife.53900

Mazor, M., Maimon-Mor, R. O., Charles, L., & Fleming, S. M. (2023). Paradoxical evidence weighting in confidence judgments for detection and discrimination. *Attention, Perception, & Psychophysics*, *85*(7), 2356‑2385. https://doi.org/10.3758/s13414-023-02710-8

Mazor, M., Moran, R., & Press, C. (2024). The role of counterfactual visibility in inference about absence. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *46*(0). https://escholarship.org/uc/item/4352x5v1

Msheik, R., Sirouet, E., Kelly, S., Faivre, N., & Pereira, M. (2025). *Leaky evidence accumulation accounts for perceptual confidence and subjective duration* (p. 2025.03.17.643717). bioRxiv. https://doi.org/10.1101/2025.03.17.643717

Plass, J., & Brang, D. (2021). Multisensory stimuli shift perceptual priors to facilitate rapid behavior. *Scientific Reports*, *11*(1), Article 1. https://doi.org/10.1038/s41598-021-02566-8

Pouget, A., Drugowitsch, J., & Kepecs, A. (2016). Confidence and certainty : Distinct probabilistic quantities for different goals. *Nature neuroscience*, *19*(3), 366‑374. https://doi.org/10.1038/nn.4240

Rach, S., Diederich, A., & Colonius, H. (2011). On quantifying multisensory interaction effects in reaction time and detection rate. *Psychological Research*, *75*(2), 77-94. https://doi.org/10.1007/s00426-010-0289-0

Schnupp, J. W. H., Dawe, K. L., & Pollack, G. L. (2005). The detection of multisensory stimuli in an orthogonal sensory space. *Experimental Brain Research*, *162*(2), 181-190. https://doi.org/10.1007/s00221-004-2136-2