Title: Judgments of agency are affected by sensory noise without recruiting metacognitive processing

Running head: Judgments of agency without metacognitive processing

Authors: Marika Constant[a,b,c], Roy Salomon[d], Elisa Filevich[a,b,c,e]

Affiliations:

[a] Humboldt-Universität zu Berlin, Faculty of Life Sciences, Department of Psychology, Unter den Linden 6, 10099 Berlin, Germany

[b] Bernstein Center for Computational Neuroscience Berlin, Philippstraße 13 Haus 6, 10115 Berlin, Germany

[c] Berlin School of Mind and Brain, Humboldt-Universität zu Berlin, Luisenstraße 56, 10115 Berlin, Germany

[d] Gonda Multidisciplinary Brain Research Center, Bar-Ilan University, Ramat Gan 5290002, Israel

[e] Center for Lifespan Psychology, Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany

Corresponding author: Marika Constant, marika.constant@gmail.com

Abstract

Judgments of agency, our sense of control over our actions and the environment, often occur in noisy conditions. We examined the computations underlying judgments of agency, in particular under the influence of sensory noise. Building on previous literature, we studied whether judgments of agency incorporate uncertainty in the same way that confidence judgments do, which would imply that the former share computational mechanisms with metacognitive judgments. In two tasks, participants rated agency, or confidence in a decision about their agency, over a virtual hand that tracked their movements, either synchronously or with a delay and either under high or low noise. We compared the predictions of two computational models to participants' ratings and found that agency ratings, unlike confidence, were best explained by a model involving no estimates of sensory noise. We propose that agency judgments reflect first-order measures of the internal signal, without involving metacognitive computations, challenging the assumed link between the two cognitive processes.

Attributing ourselves agency, or causation of our actions and their outcomes, is central to our experience of moving intentionally. Previous research has suggested that agency depends on a comparison between the predicted and observed consequences of our actions, resulting in a prediction error signal if they do not match. Under this widely accepted comparator model of agency, it is the prediction error signal that leads to our feeling of agency (FoA), which we assess in order to make judgments of agency (JoAs) (Carruthers, 2012; Frith et al., 2000; Haggard, 2017). There are, however, several sources of noise to these signals, and agency processing, much like other perceptual processing, occurs under varying degrees of uncertainty. Noisy agency signals are suggested to play a role in clinical cases involving striking disruptions to agency, such as in schizophrenia (Corlett et al., 2010; Fletcher & Frith, 2009; Moore & Fletcher, 2012; Robinson et al., 2016), so it is critical to understand precisely how agency changes with uncertainty, in both healthy and clinical populations. While previous literature has investigated the role of uncertainty in FoAs (Legaspi & Toyoizumi, 2019; Moore & Haggard, 2008), it remains unclear how this affects explicit JoAs. JoAs are often considered to be second-order, metacognitive reports about otherwise first-order agency signals (Metcalfe et al., 2012; Metcalfe & Greene, 2007; Miele et al., 2011; Wenke et al., 2010). If this assumption were valid at the computational level, JoAs should then monitor noise similarly to metacognitive judgments such as confidence. On the other hand, the assumed relationship between JoAs and metacognition might only be based on a conceptual level, with JoAs conforming to a definition of metacognition as cognition about our own cognition (Flavell, 1978; Fleming et al., 2012). Here, we focussed on the relationship between JoAs and confidence to determine whether it is valid not only at the broad conceptual level, but also holds at the level of uncertainty monitoring computations.

In metacognition research, metacognitive ability is often operationalized as the precision of confidence ratings following discrimination responses (Fleming & Lau, 2014). Under this narrower operationalization, metacognitive confidence ratings have been shown to monitor the uncertainty in the perceived internal signal (Gardelle & Mamassian, 2015; Navajas et al., 2017; Rausch et al., 2018; Sanders et al., 2016; Spence et al., 2016). Thus, here we commit to a definition of metacognition as a process that involves second-order uncertainty monitoring computations. We propose that for agency judgments to be metacognitive in a computational sense, they should monitor the noise in the perceived prediction error signal, incorporating a

second-order judgment of the uncertainty of one's agency processing. Some existing models of agency have suggested a role of noise in the comparator model by showing that the comparator signal is the result of a cue integration process, in which the motor and sensory information are integrated and weighted by their reliability (Moore & Fletcher, 2012; Moore & Haggard, 2008). In line with these models, participants' FoA, as measured by an implicit temporal binding effect, has been found to rely less on cues that are noisy or imprecise, and more on more informative cues (Moore et al., 2009; Wolpe et al., 2013). Importantly, this work, which has focussed on the precision of the comparator signal itself by determining how information gets optimally integrated to form that signal, has not addressed whether agency judgments involve a second-order monitoring of that precision. Further, one recent study suggested a role for sensory uncertainty in FoAs by proposing a measure of confidence in one's causal estimate (CCE) as the computation underlying them (Legaspi & Toyoizumi, 2019). However, this model does not explain how the suggested precision-dependent FoAs relate to the level of explicit JoAs. It is not clear that explicit JoAs would incorporate noise in the same way as lower level FoAs, as they are generally considered to be at a higher level of the processing hierarchy (Gallagher, 2007; Haggard & Tsakiris, 2009; Sato, 2009; Synofzik et al., 2008).

Here, we expanded on previous work and directly investigated the question of whether JoAs are metacognitive in a computational sense. We did this by setting up a two-criterion test. The first criterion was for sensory noise to influence JoAs, beyond altering their variance across trials. This would suggest that the reliability of the signal was factored into each rating, against the simpler alternative that the rating was made on the basis of a linear readout of the less reliable signal. We examined this by assessing the effects of noise and delay on explicit agency ratings, using a sensory noise manipulation orthogonal to the delay. This first criterion formed our pre-registered hypothesis and was necessary, but not sufficient, for JoAs to be considered metacognitive. It therefore served as a prerequisite for the second criterion: Agency and confidence judgments should show similar sensitivity to internal estimates of the sensory noise, suggesting the involvement of second-order uncertainty monitoring following the same computational principles. We assessed this by contrasting two computational models against distributions of JoA data, one that would reflect metacognitive monitoring of noise and an alternative model that would not. We found that JoAs satisfied the first, but not the second criterion: Sensory noise did indeed influence JoAs, but this influence did not reflect any

second-order noise monitoring, suggesting that JoAs may not be metacognitive in the computational sense.

Results

Each participant completed two tasks: A confidence-rating task, consisting of a two-interval forced choice (2IFC) followed by a confidence rating on a scale from 1 to 6; and an agency-rating task consisting of a JoA on an equivalent scale. Both tasks used the same basic stimuli, namely, a movement of participants' index finger tracked by a LEAP Motion infrared tracker and displayed on the screen as a virtual hand movement either in synchrony or with small temporal delays. In both tasks, we manipulated sensory noise in the same way, by changing the illumination of the scene. We created two conditions (low and high sensory noise) by displaying the virtual hand under bright, high contrast illumination or under dim, low contrast illumination respectively (Fig. 1A).

Confirmatory Analyses

Confidence-rating task. On each trial of the confidence task, participants were cued to make two consecutive movements of their right index finger, with their hand out of sight. For only one of the two intervals, we added a temporal delay to the virtual hand shown on the screen (in the other interval, the virtual hand was displayed to match the participant's hand movement in real time). Participants then discriminated in which interval they felt more agency over the movement of the virtual hand, and rated confidence in their response (Fig. 1B). We assumed that participants compared their degree of control over the virtual hand in the two movements to solve the task, and rated confidence in this comparison. This paradigm brings agency into a standard framework for studying metacognition (Wang et al., 2020). Importantly, under this operationalization, we can define correct responses to the 2IFC discrimination task as those where participants report that they felt more agency for the stimulus without any added delay, allowing us to quantify discrimination accuracy. If the illumination manipulation served to increase sensory noise in the intended way, we expected lower discrimination accuracy under high sensory noise compared to low noise (Macmillan & Creelman, 1991). Further, based on previous work using similar confidence paradigms (Bang & Fleming, 2018; Gardelle & Mamassian, 2015; Spence et al., 2016), and a normative model of confidence (Sanders et al.,

2016), we predicted an interaction between sensory noise and accuracy on confidence, in particular with confidence decreasing in high noise following correct trials and increasing in high noise following incorrect trials. To test the effect of the illumination manipulation, we first built a logistic regression model on response accuracy, including sensory noise as a fixed effect, and by-participant random intercepts (see Table 1 for the explicit model syntax). As expected, we found a main effect of Noise, revealing significantly lower accuracy in the high-noise compared to the low-noise condition (Mdiff = 10%, SE = 1.4%, $\chi^2(1) = 97.60$, $p < 0.001$, $BF_{10} = 1.78 \times 10^{20}$, OR = 1.55, 95% CI [1.42, 1.70]; Fig. 1C). Then, we built a linear mixed-effects model on confidence to test the second prediction (an interaction effect between sensory noise and response accuracy). The model included the interaction between response accuracy and noise level and each factor as fixed effects, as well as by-participant random intercepts. We also included response identity (first or second interval) as a fixed effect, as presentation order could have biased confidence ratings (Jamieson & Petrusic, 1975; Yeshurun et al., 2008). In line with our predictions, we found a significant interaction between Noise and Response Accuracy on confidence, $F(1,8858) = 14.43$, $p < 0.001$, $BF_{10} = 2.19$, $\eta^2_p = 0.0016$ (Fig. 1D), with a stronger difference in confidence between correct and incorrect trials under low noise ($Mdiff_{Correct-Incorrect} = 0.58$, SE = 0.044) compared to high noise ($Mdiff_{Correct-Incorrect} = 0.34$, SE = 0.042). In addition to the interaction effect, we found that confidence following incorrect decisions was lower in the high-noise compared to the low-noise condition. Although we expected confidence following incorrect decisions to increase under high noise, the 'double-increase' confidence pattern seen here has also been shown in the literature (Adler & Ma, 2018; Rausch et al., 2018; Rausch & Zehetleitner, 2019). Finally, we found a significant main effect of Response, $F(1,8872) = 82.64$, $p < 0.001$, $BF_{10} = 5.71 \times 10^{14}$, $\eta^2_p = 0.0092$, with pairwise comparisons revealing significantly higher confidence ratings when participants reported feeling more agency over the stimulus in the second interval, compared to the first (Mdiff = -0.27, SE = 0.030), $t(8872) = -9.09$, $p < 0.001$. These results were also all confirmed by repeating this analysis with ordinal models (Supplementary Information). Together, these results suggest that the illumination manipulation affected sensory noise as we intended, and influenced metacognitive confidence judgments as previous studies have shown. Having validated our experimental manipulation, we then went on to analyze the effect of sensory noise on JoAs.

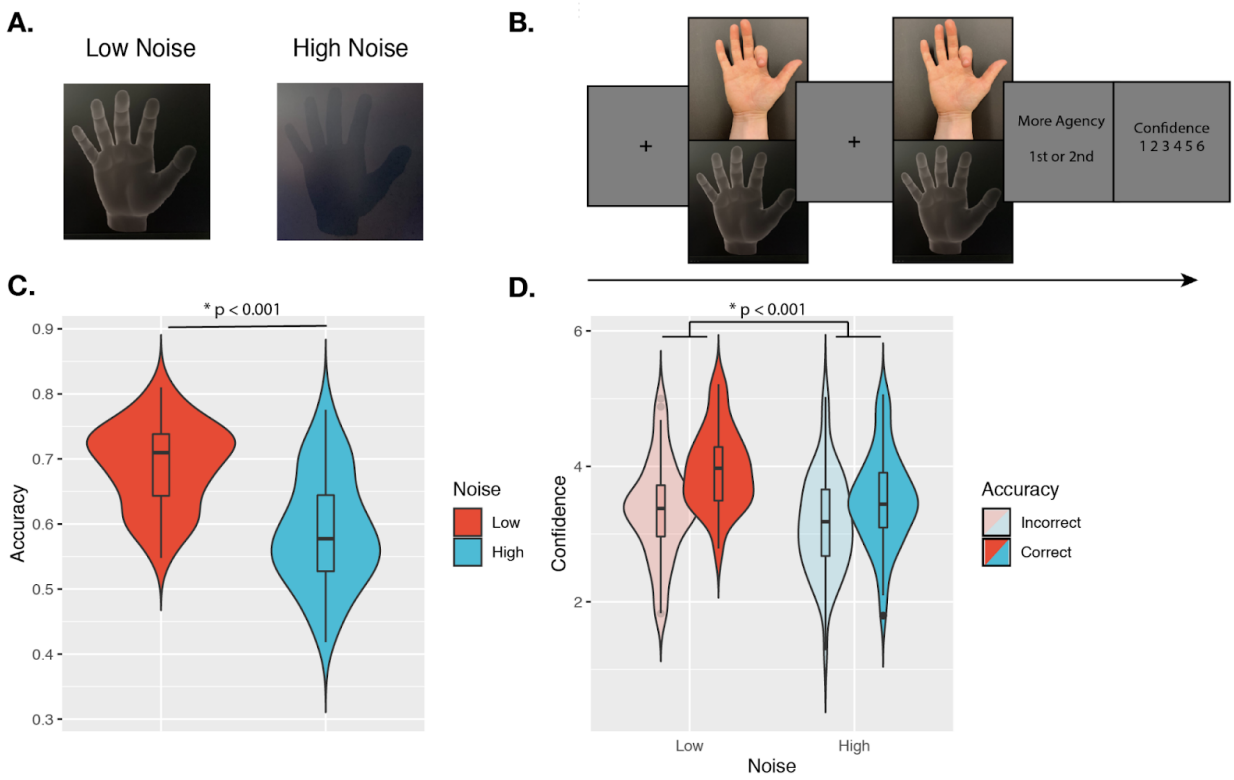


Figure 1: Confidence Task. (A.) Sensory noise conditions. Sensory noise was manipulated by changing the illumination, with high sensory noise captured by low contrast, dark illumination, and low sensory noise captured by high contrast, bright illumination. **(B.) Experimental paradigm.** Cued by the offset of the fixation cross, participants made two consecutive finger movements on each trial, with their hands out of sight. On the screen, participants saw either a virtual hand moving in synchrony with their own, or with an additional temporal delay. Participants first discriminated which movement they felt more agency over, and then rated their confidence in their own response. **(C.) Discrimination accuracy.** In line with the intended effect of the sensory noise manipulation, accuracy was significantly higher for low- vs. high-noise conditions. **(D.) Mean confidence ratings.** We found a significant interaction effect between Response Accuracy and Noise on Confidence. Violin plots capture the kernel probability density and boxplots show the median, interquartile range (IQR) with hinges showing the first and third quartiles, and vertical whiskers stretching to most extreme data point within 1.5*IQR from the hinges. Outliers are plotted as black or grey dots.

Agency-rating task. On each trial of the agency-rating task, participants made a single movement of their index finger and watched the virtual hand model either move synchronously with their movement (25% of trials) or with a delay of either 70, 100, or 200 ms. After every movement, participants rated their agency on a scale from 1 to 6 (Fig. 2A). By adding sensory noise to the perceived sensory outcome of the movement (namely, the virtual hand movement), we added noise to the comparator signal that participants assessed with their JoAs (Fig. 2B). We

then formulated a two-criterion test to assess whether JoAs are computationally metacognitive. The first criterion was for sensory noise to influence agency ratings beyond just increasing their variability, and hence for JoAs to reflect not only a readout of the comparator signal but also an indication of the signal's precision (Fig. 2B). For this criterion to be met, mean JoAs should depend on both delay and sensory noise. Alternatively, if the mean JoA per delay does not depend on the sensory noise level, this would indicate that JoAs are simply a first-order report of the perceived comparator signal, with mean JoA reflecting the mean of the comparator signal distribution (Fig. 2B). The second criterion of our test, if the first criterion was met, was for agency ratings to involve underlying metacognitive computations such as those involved in confidence, in particular, the second-order monitoring of sensory noise. To test this, we compared two computational models built based on the results of the first criterion, a Bayesian-agency model that involves metacognitive processing, and an alternative Rescaling model that does not.

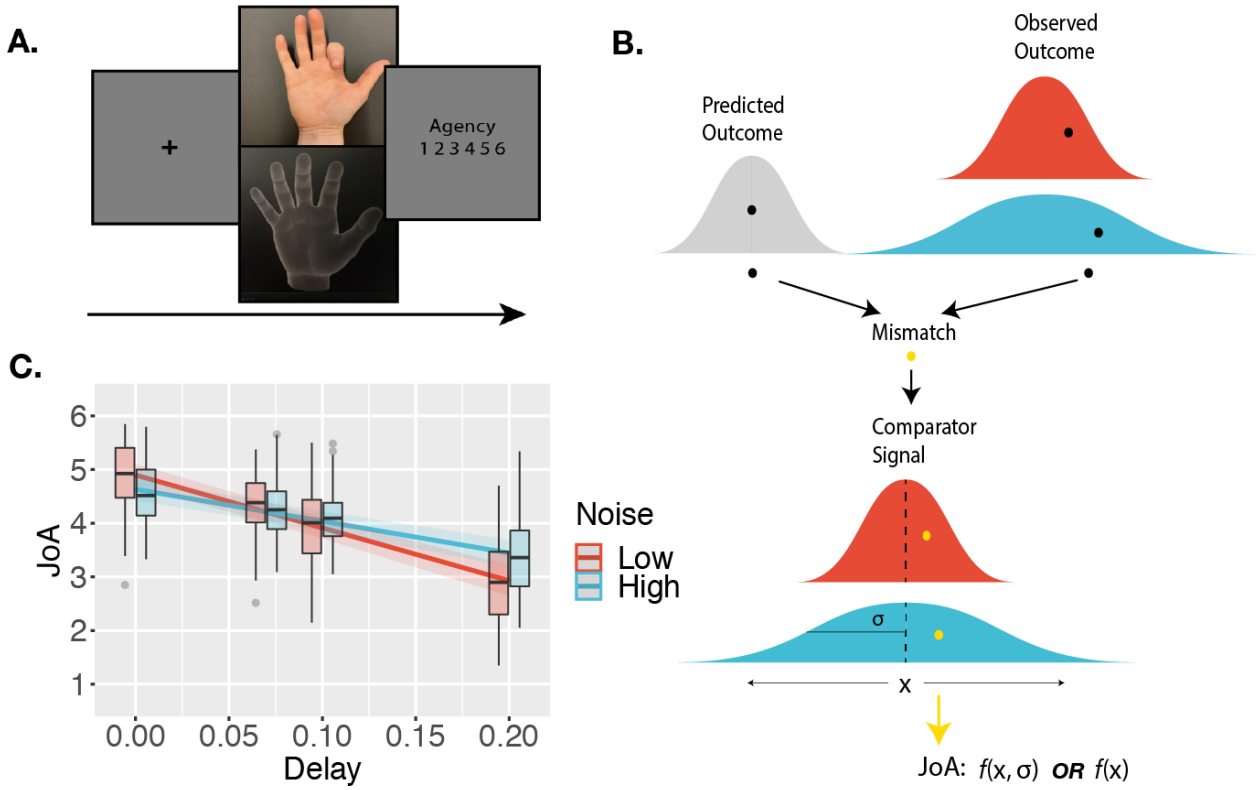


Figure 2: Agency-rating Task. (A.) Experimental paradigm. Cued by the offset of the fixation cross, participants made one finger movement on each trial, with their hand out of sight. On the screen, participants saw either a virtual hand moving in synchrony with their own, or with an additional delay/temporal lag. Participants then made a judgment of the degree of agency they experienced. **(B.) Sketch of noise dependency prediction.** Our high noise condition adds sensory noise to the observed outcome, and hence to the comparator signal, computed as the mismatch between predicted and observed outcomes. Points reflect the perceived signals on single trials. If the JoA is a readout of the comparator signal on each trial, though the variance would increase with noise, the mean JoA will not depend on noise, but will reflect the mean of the comparator signal distribution. Alternatively, if JoAs monitor the noisiness of the comparator signal, mean JoA will depend on noise. In other words, JoAs would be a function not only of a noisier signal (x), but of the noisier signal and the noise itself (σ). **(C.) Interaction effect result.** Predicted JoA across delays and noise conditions from linear mixed-effects model results. 95% confidence intervals shown. Boxplots reflect subjectwise mean JoAs per noise level and delay. They show the median, interquartile range (IQR) with hinges showing the first and third quartiles, and vertical whiskers stretching to most extreme data point within 1.5*IQR from the hinges. Outliers are plotted as grey dots.

Behavioural Results. While the results of the confidence task confirmed that the illumination manipulation affected sensory noise overall as intended, we were interested in examining precisely how JoAs responded to sensory noise, and hence required that the high-noise condition actually increased sensory noise for all participants included in the analysis

of the agency task. We therefore excluded from the following analyses any participants for whom discrimination accuracy in the high-noise condition was not lower than in the low-noise condition in the confidence task. We note that all of the results described below remain largely the same, and the conclusions unchanged, when we included the data from all participants in the analyses. We predicted that if sensory noise affected JoAs similarly to metacognitive processes, we would observe a significant interaction between Noise and Delay on JoA (Fig. 2B). This would be the first of our two criteria. We investigated this using a linear mixed-effects model on JoAs that included the interaction between noise level and delay as fixed effects, and allowed for by-participant random effects of the interaction, and random intercepts (Table 1). We found a significant interaction effect between Noise and Delay, $F(1,52) = 61.16$, $p < 0.001$, $BF_{10} = 3.78 \times 10^6$, $\eta^2_p = 0.54$, 95% CI [0.35, 0.67], with a less extreme negative slope across delay values in the high-noise condition ($\beta_{High} = -5.93$, SE = 0.69), compared to low-noise ($\beta_{Low} = -9.84$, SE = 0.77) (Fig. 2C), suggesting that JoAs met our first criterion. We also found a significant main effect of Delay, $F(1,39) = 132.05$, $p < 0.001$, $BF_{10} = 14486.52$, $\eta^2_p = 0.77$, 95% CI [0.64, 0.85], replicating previous findings that showed increasing delays of the virtual hand movement to lead to lower JoAs (Krugwasser et al., 2019; Stern et al., 2020). We found this effect of JoAs decreasing with delay for the majority of participants (37 out of 40) included in this analysis in both conditions, indicating that participants were able to make meaningful ratings, even in the high-noise condition. We also repeated these analyses with ordinal models, which confirmed our results (Supplementary Information).

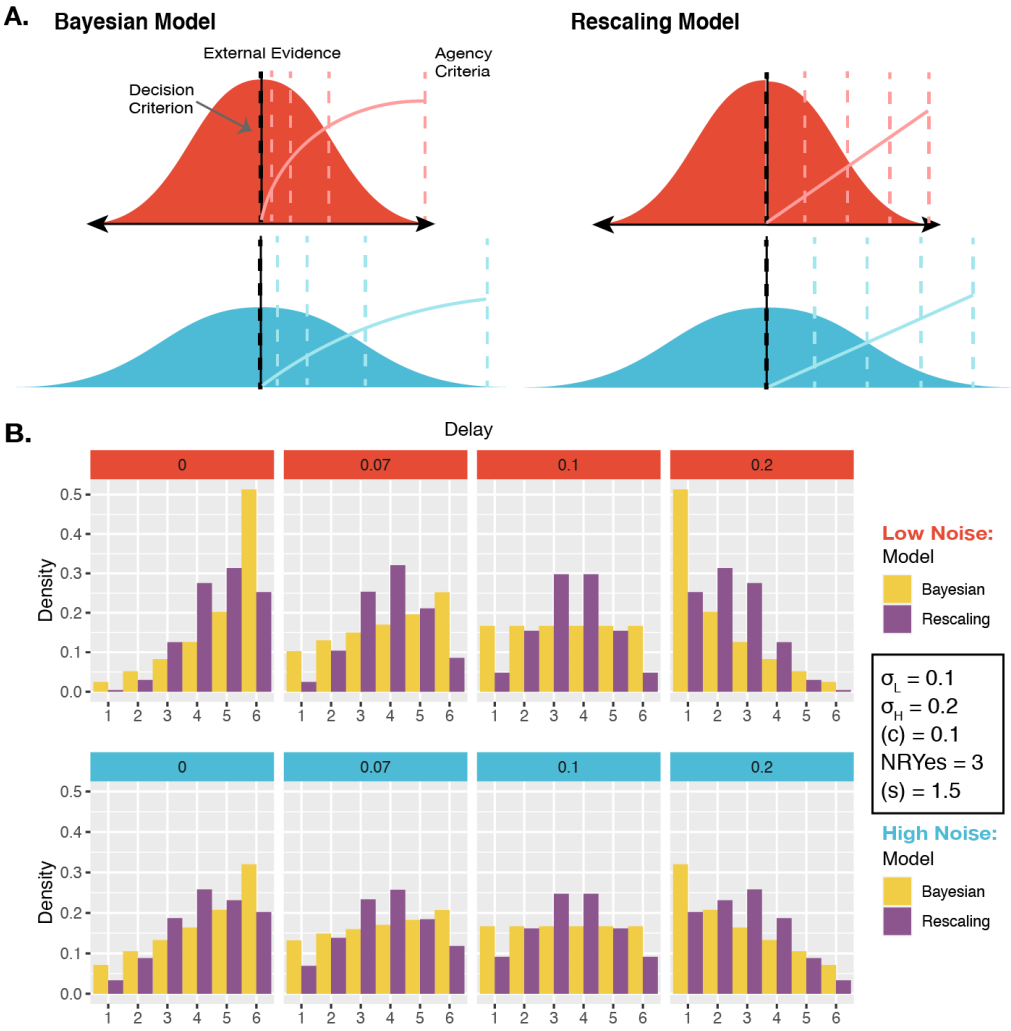


Figure 3: Models and Predictions. (A.) Two models of JoAs. In the **Bayesian model of agency**, JoA reflects the posterior probability of the agency detection decision being correct, given the choice and internal evidence. The agency criteria are spaced linearly in probability space, so their positions on the internal signal space change with noise level. The solid coloured lines show agency as a function of internal signal strength. In the **Rescaling model of agency,** JoA reflects a first-order estimate of delay compared to the criterion, not based on noise. However, the agency criteria are spread evenly across the range of signals within each noise level, such that they interact with noise level only due to a rescaling of ratings. The solid coloured lines show agency as a function of internal signal strength, and this function is linear but still interacts with noise. **(B.) Model predictions.** Predictions were based on simulations run with representative parameters, shown. Predicted distributions of JoAs per delay and noise condition are shown as densities.

Exploratory Analyses

Computational Modeling. We found in our behavioural analysis that the mean JoA depended on both delay and noise, meeting our first criterion for JoAs being metacognitive. This

allowed us to move on to our second test-criterion and investigate whether there are strictly metacognitive computations underlying agency judgments. In order to test our second criterion, namely, whether JoAs can be explained by the same computations as confidence, we compared two possible models of agency ratings (the Bayesian-agency model and the Rescaling model, Fig. 3A), that differed in their predicted distributions of JoAs across noise conditions and delays (Fig. 3B). Both models could in principle account for the observed interaction effect between noise and delay on JoA, satisfying the first criterion, but only the Bayesian model included a metacognitive assessment of one's own sensory noise. The Bayesian-agency model assumed that agency ratings behave like confidence as described by Bayesian-confidence models, namely as the posterior probability that a decision is correct, given the strength of the internal evidence and the decision (Navajas et al., 2017; Sanders et al., 2016). The computation of this probability requires the observer to have second-order access to estimate their own sensory noise (Fig. 3A). As an alternative, we considered the Rescaling model, which parallels the Bayesian one except that ratings are based on first-order point-estimates of evidence and therefore do not involve metacognitive estimates of sensory noise being factored into ratings (Fig. 3A). This Rescaling model accounts for the observed relationship between noise and JoAs by considering that participants might rescale their ratings based on the noise condition. In practical terms, this implies that participants treated the conditions independently during the task, judging agency in low-noise trials relative to one another, and high-noise trials relative to one another. Although our design aimed to prevent this by interleaving the conditions, it is still possible that participants did this to some extent, as our noise manipulation was visually very apparent. Critically, this rescaling is achieved without making any estimates of the sensory noise of the conditions. This would mean that in the Rescaling model a JoA of '6' could have referred to different strengths of agency experience between the two noise conditions, as it is rescaled to reflect the highest strength agency experience of only the trials of that condition. In the Bayesian model, on the other hand, a JoA of '6' would always reflect the same perceived level of certainty about the agency decision and hence the same agency experience, since JoAs combine the evidence strength and sensory noise.

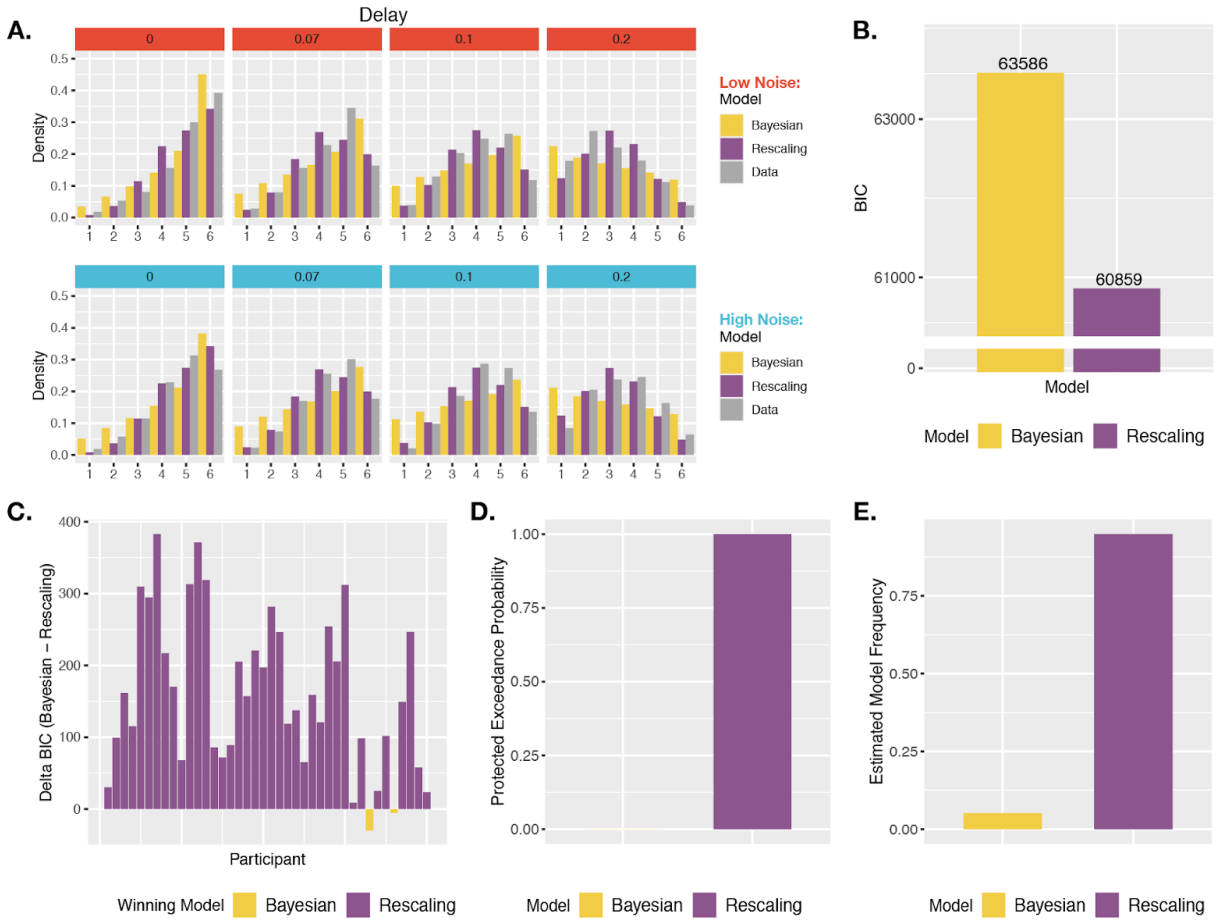


Figure 4: Agency Model Fits and Results. (A.) Model fits. Simulated probability of each JoA for a given delay and noise level, given best fitting parameters from the MLE analysis for Bayesian and Rescaling models. These are portrayed as densities across the JoAs. Data distributions are shown in shaded grey, also portrayed as densities across discrete ratings. **(B.) Group BIC Results.** BIC comparison between Bayesian and Rescaling models on pooled JoA data. **(C.) Subject-wise BIC Results.** Difference in BIC (Bayesian - Rescaling) for each participant, with negative values indicating that the Bayesian model fit better, and positive values indicating that the Rescaling model fit better. **(D.) Protected Exceedance Probabilities.** Protected exceedance probabilities of the Bayesian and Rescaling models. **(E.) Estimated Model Frequencies.** Predicted model frequencies estimated from the exceedance probability analysis.

We compared the two models in their ability to account for the distributions of JoAs per delay and noise level (Fig. 4A) at the group level (with pooled data) and at the single-participant level. We also measured protected exceedance probabilities (PEPs) to assess the probability that each model was the most frequently occurring one in the population, adjusted for the ability to reject the null hypothesis (that the models are equally likely in the population). To compare the

models, we first found the best fitting parameter values for each model — low noise σ_L, high noise σ_H, decision criterion (c), and the mapping parameter, or the number of ratings to be considered as 'Yes' responses (NRYes), for both models, as well as the scale range parameter (s) for the Rescaling model — using maximum likelihood estimation (MLE), and then performed a Bayesian-information criterion (BIC) comparison. Following standard recommendations (Raftery, 1995), we required a minimum BIC difference of 2 in order to consider one model a better explanation of the data than the other. To perform the Bayesian model selection and calculate the PEPs, we used the 'bmsR' package (Lisi, 2021) with model evidence computed from the Akaike weights (Rigoux et al., 2014; Stephan et al., 2009; Wagenmakers & Farrell, 2004). We also performed a model recovery analysis across the set of winning parameter values of each model across participants. This confirmed that the models were distinguishable from one another over the entire relevant parameter space, with the Bayesian model being correctly recovered 97.00% of the time, and the Rescaling model being correctly recovered 99.88% of the time. The full model recovery analysis can be found in Supplementary Information.

Against the notion that agency ratings and confidence arise from analogous uncertainty monitoring computations, the group-level analysis revealed that the Rescaling model could better explain the JoA data ($\Delta BIC_{Bayes\text{-}Rescaling}$ = 2728, Fig. 4B). The best fitting parameters for this Rescaling model were σ_L = 0.16, σ_H = 0.16, (c) = 0.16, NRYes = 3, and (s) = 1.11. In comparison, the best fitting parameters for the Bayesian model were σ_L = 0.19, σ_H = 0.24, (c) = 0.16, and NRYes = 3. The predicted densities of each rating per delay and noise level for each model's best fitting parameters can be seen in Fig. 4A. Fitting the two models to each participant revealed results consistent with the group-level analysis: The Rescaling model could better explain the data for 38 out of 40 participants, whereas the Bayesian model provided a better fit for only 2 (Fig. 4C). Also in line with this, the PEPs indicated that the Rescaling model occurs most frequently in the population (Fig. 4D), with the predicted frequencies shown in Fig. 4E.

We then performed the same model comparison on confidence ratings from the confidence task in order to confirm metacognitive computations underlying confidence using this modeling approach, for comparison to our agency-rating results. As expected, the Bayesian model could better explain confidence ratings ($\Delta BIC_{Rescaling\text{-}Bayes}$ = 1121, Fig. 5A), suggesting confidence to involve metacognitive computations, in contrast to JoAs. The PEPs also suggested the Bayesian model to be the most frequently occurring model in the population (Fig. 5B), with

estimated frequencies shown in Fig. 5C. The subject-wise BIC comparison revealed the Bayesian model to provide a better fit in 24 out of 40 participants, the Rescaling model to provide a better fit in 13, and a BIC difference of less than 2 (suggesting neither model being a conclusively better fit) in 3 (Fig. 5D).

Metacognitive Ability. In order to further ensure that participants' confidence ratings reflected metacognitive processing, especially given relatively low accuracy under high noise, we analysed their metacognitive ability in both noise conditions. We did this by computing metacognitive efficiency (M-Ratio), which accounts for differences in first-order task performance (Maniscalco & Lau, 2012), using the HMeta-d′ toolbox (Fleming, 2017) for all participants. This revealed above-chance metacognitive efficiency (M-Ratio > 0) in both noise conditions. Importantly, we did not aim to compare metacognitive ability between conditions, but instead only confirmed that participants showed above-chance metacognitive performance in both conditions, suggesting in turn that confidence ratings were meaningful in both. Interestingly, we also found that metacognitive efficiency was nearly indistinguishable between conditions ($\text{M-Ratio}_{\text{Low Noise}} = 0.73$, $\text{M-Ratio}_{\text{High Noise}} = 0.74$, Fig. 5E), once differences in first-order accuracy were accounted for. These results were further confirmed by an analysis of metacognitive sensitivity using logistic regressions (Supplementary Information).

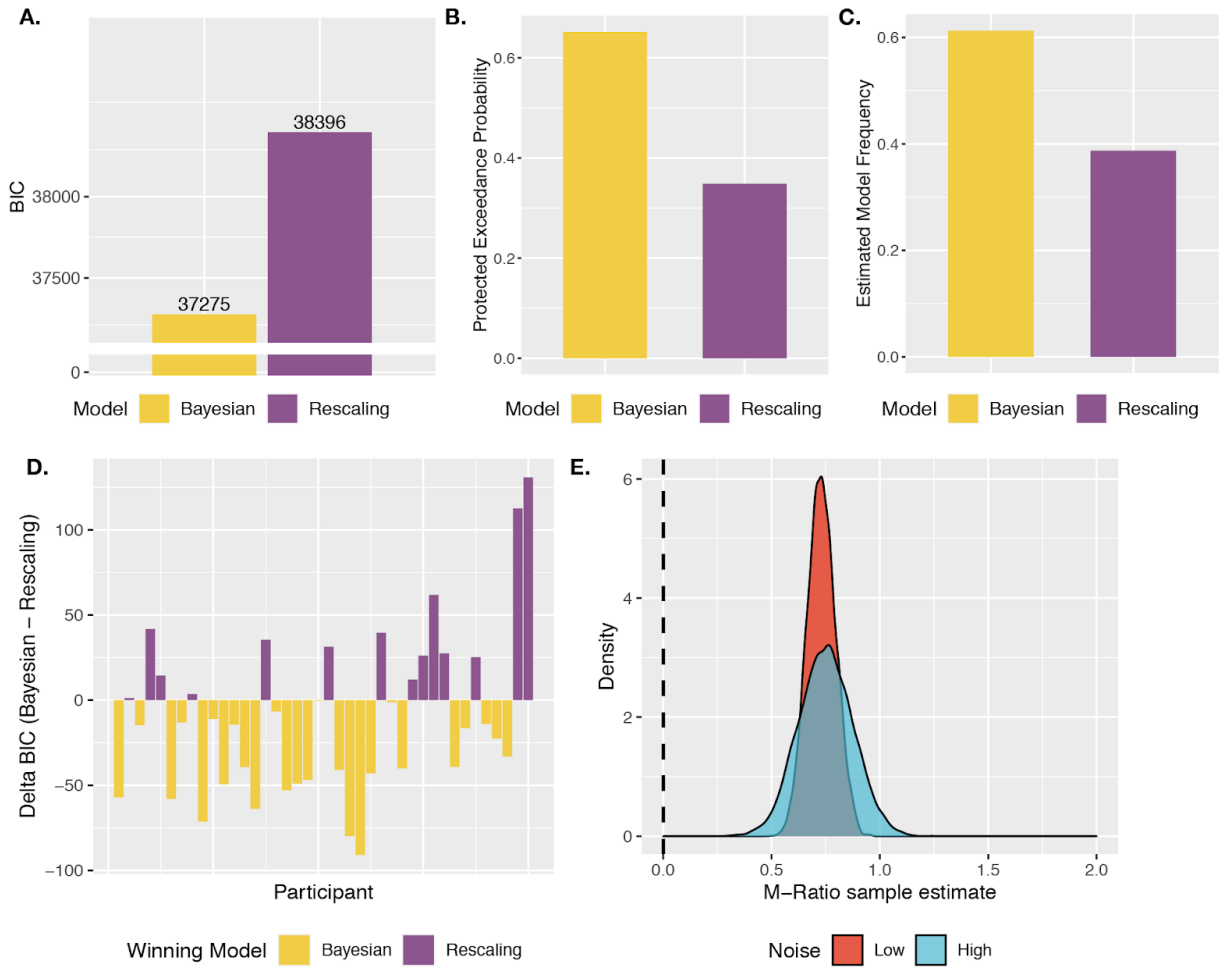


Figure 5. Confidence and Metacognitive Performance Results. (A.) Group BIC Results. BIC comparison between the Bayesian and Rescaling models on pooled confidence rating data. **(B.) Protected Exceedance Probabilities.** Protected exceedance probabilities of the Bayesian and Rescaling models. **(C.) Estimated Model Frequencies.** Predicted model frequencies estimated from the exceedance probability analysis. **(D.) Subject-wise BIC Results.** Difference in BIC (Bayesian - Rescaling) for each participant, with negative values indicating that the Bayesian model fit better, and positive values indicating that the Rescaling model fit better. **(E.) M-Ratio Estimates.** Metacognitive efficiency for each noise level, estimated using the HMetad′ toolbox. The dashed vertical line indicates chance-level metacognitive efficiency.

Discussion

Despite previous research investigating the effect of noise on implicit proxies of agency such as temporal binding effects, it has remained unclear how explicit agency incorporates uncertainty. Meanwhile, agency judgments are often assumed to be metacognitive without explicitly defining or testing this relationship. Here, we aimed to dissect this assumption by

examining whether JoAs can be said to be metacognitive only at a broad conceptual level, or whether they also involve the same uncertainty monitoring computations as confidence judgments, widely thought to be metacognitive. That is, we evaluated whether JoAs involve second-order noise estimation or are rather grounded in first-order sensorimotor signal integration. By combining two tasks, we brought JoAs into a standard metacognitive framework and compared them to confidence ratings following a 2IFC decision. By examining how discrimination accuracy and confidence changed with sensory noise in the confidence-rating task, we first confirmed that the sensory noise manipulation had the intended effect.

The effects of sensory noise on confidence allowed us to consider whether JoAs in the agency-rating task responded to sensory noise in a computationally analogous way. We reasoned that, if this were the case, JoAs would satisfy two criteria: First, they would depend on the precision of the comparator information, reflecting more than just a readout of the perceived signal. Second, this dependence on noise would reflect underlying metacognitive computations such as those involved in confidence, in particular, second-order estimates of one's own sensory noise. The JoAs satisfied the first of the test-criteria. We found that noise did indeed influence mean JoA across delays, indicating that the noise condition is factored into JoAs. However, because the noise manipulation changed the display in a visually obvious way, this information about the condition could have influenced judgments in a way that did not reflect participants making metacognitive estimates of the noise of their own processing. The second test-criterion investigated precisely this possibility.

To assess the second criterion, we compared two computational models. As a prerequisite, both models satisfied the first test-criterion. We contrasted a Bayesian agency-model, which included metacognitive noise estimates, with a Rescaling model that did not imply metacognitive processing. We tested these models in their ability to fit participants' JoAs as well as confidence judgments, to understand the computations underlying both. In the case of confidence, the model comparison revealed that participants' judgments were best explained by the Bayesian model, confirming the metacognitive computations underlying them (Kepecs & Mainen, 2012; Meyniel et al., 2015; Pouget et al., 2016; Sanders et al., 2016). In striking contrast, this model comparison against participants' agency ratings revealed that JoAs were better explained by the Rescaling model as compared to the Bayesian agency-model, even despite the former including one additional free parameter. The Rescaling model accounted for

the observed behavioural relationship between JoA and noise by assuming that participants compared trials only to other trials within the same condition, and set condition-specific maxima of their rating scales accordingly. It would be an interesting direction for future work to test how JoAs depend on sensory noise under a noise manipulation that is not easily detectable, to investigate if the behavioural relationship we observe between noise and JoAs is limited to cases in which participants can treat the manipulation conditions as independent contexts, as our models suggest.

Taken together, our results suggest that while JoAs can be influenced by sensory noise, this influence is not indicative of metacognitive processing, and JoAs may better reflect first-order assessments of agency signals. We therefore argue that greater care should be taken when discussing agency within a metacognitive context, as the assumptions made about agency judgments being metacognitive do not hold on a computational level. Although this work used confidence as a benchmark for metacognitive processing, the computation of interest is second-order monitoring of the precision of one's processing, which has become the narrower focus of recent metacognition work (Fleming & Lau, 2014). While JoAs may still satisfy broad definitions of metacognition, our results suggest that they may not satisfy this narrower definition that is associated with a concrete computational view. In this sense, these results may help specify and clarify the assumed relationships between explicit agency judgments and metacognition. At the same time, they may add to our understanding of how JoAs respond to uncertainty — similarly to first-order perceptual judgments — which is critical for interpreting agency reports, especially in certain clinical cases involving agency disruptions.

Importantly, we also found that participants can make metacognitive confidence judgments *about* agency decisions, but that subjective agency ratings do not share these computations, despite the same basic agency task and noise manipulation. The use of 2IFC agency tasks with confidence has recently been proposed as a promising step towards a more reliable and complete investigation of agency processing, both in healthy and clinical populations (Wang et al., 2020). Here, using a virtual hand we extended this approach into a more proximal form of embodied agency, closer to agency over the body itself (M.S. Christensen & Grünbaum, 2018; Dogge et al., 2019; Stern et al., 2020; Wen, 2019), and provide an initial step in demonstrating that participants can meaningfully monitor the accuracy of these agency decisions. We suggest that confidence judgments about agency should be considered as the

metacognitive level of an agency processing hierarchy, with agency judgments as explicit first-order judgments. This also brings agency in line with recent motor metacognition research that considers agency-like judgments such as decisions of which trajectory was caused by one's movement to be the first-order motor judgments, followed by metacognitive confidence ratings (Arbuzova et al., 2021).

Although we suggest that JoAs do not imply metacognitive uncertainty monitoring, the dependence of agency ratings on the noise condition should not be overlooked and may be highly relevant both for future experimental design and in the interpretation of explicit agency reports. This finding is in line with multifactorial accounts of JoAs as involving a variety of both internal and external cues (Synofzik et al., 2008), and with the expanding empirical work investigating a range of contextual effects on agency (Minohara et al., 2016; Wen, 2019). Our work also fits within cue integration theories of agency (Moore & Fletcher, 2012; Synofzik et al., 2013), with the delay information being weighted less heavily when the signal is made less precise. Further, these results are relevant to empirical work examining cue integration in agency, as they suggest that having a perceivable manipulation such as reduced visibility in order to add noise to feedback cues may itself act as an additional factor influencing agency judgments, which should be accounted for in design and analysis.

These findings also complement recent work that has aimed to find computational models of agency, but has focussed on low-level FoA and implicit measures such as temporal binding effects (Legaspi & Toyoizumi, 2019). Here, we bring explicit agency judgments into a Bayesian and SDT framework, implementing formal computational models that could be used to further assess computations underlying different JoAs. Our findings support the suggestions of previous work that, while Bayesian confidence computations may underlie pre-reflexive FoA, explicit JoAs reflect a different computational mechanism, and factor in different contextual information (Legaspi & Toyoizumi, 2019; Wen, 2019).

Taken into the context of two-step models of agency (Synofzik et al., 2008), our results suggest that sensory noise information may influence lower level, perceptual agency signals by making them more variable, and that these more variable signals then feed into higher level JoAs, but that estimates of the noise do not contribute as an additional cue to inform the JoAs. Hence, higher order agency states remain naive with regard to how noisy the lower level signals are. This could illuminate agency processing deficits in clinical cases which might involve

perceptual agency cues becoming very noisy due to low level processing deficits. If explicit JoAs do not consider this uncertainty (as our findings indicate), this could lead to extreme agency reports despite unreliable evidence and inaccurate agency inferences. This is in line with work on agency misattributions in schizophrenia suggesting that they may be due to particularly noisy low level agency information, possibly due to dysregulated neurotransmitter activity (Corlett et al., 2010; Fletcher & Frith, 2009; Moore & Fletcher, 2012; Robinson et al., 2016).

Our combination of tasks allowed us to consider JoAs against the benchmark of metacognitive confidence judgments, while keeping the basic stimuli and noise manipulations the same in both cases. Despite this, the analysis of JoAs is still limited by the type of ratings we collected, and the manipulations used in the task. Our noise manipulation, for example, was limited to adding noise to the perceived outcome signals, rather than motor or somatosensory cues. Our approach of adding sensory noise orthogonally to the degree of control could be expanded to add noise to other agency signals, and it is possible that JoAs monitor these sources of noise differently. Beyond the sensory noise manipulation, our results should also be considered in the context of the particular agency rating scale we used, as this will constrain participants' rating behaviour. Similar agency scales are used in other agency research (Dewey & Knoblich, 2014; Imaizumi & Tanno, 2019; Kawabe et al., 2013; Metcalfe & Greene, 2007; Miele et al., 2011; Sato & Yasuda, 2005; Stern et al., 2020; Voss et al., 2017; Wen et al., 2015), so the results presented here are relevant to understand the computations of agency ratings discussed in the existing literature, to the extent that they are based on similar experimental operationalizations. In light of this, although our agency-rating task differs from our confidence task and does not include any postdecisional component, which could be argued to underlie the differences in metacognitive processing, our results still suggest that the type of agency judgments measured in the literature do not indicate metacognitive processing, whether or not this is due to a lack of postdecisional ratings. On the basis of these results, we concur with recent work (Wang et al., 2020) that suggests that a 2IFC task on agency followed by a confidence judgment may be more adequate to measure a metacognitive component of agency processing. Also, while it is true that the rating scale we used here presupposes that agency is graded, the JoA models assessed here actually allowed for the possibility of a binary agency detection threshold, with scaling based on certainty level. Though our results show confidence to be a poor explanation of the scaling of agency ratings, future work could apply a similar modeling

approach in order to investigate other possibilities, and to explore other types of agency judgments. Future work could also further investigate another simplifying assumption made here in our models, namely that trials in which, due to internal noise, participants experienced a negative delay, would be associated with strong evidence for agency. While this is a reasonable assumption in our task, where participants knew that the virtual hand tracked their hand movement and likely dismissed the possibility that the virtual hand moved prior to them, it may need to be adapted to fit other, more ecologically valid cases of agency processing.

In conclusion, here we brought agency judgments into a metacognitive framework in order to directly assess whether JoAs are metacognitive at the computational level. The results suggest that agency ratings can be influenced by sensory noise, but that this effect is best considered as a contextual cue that can affect participants' scaling of agency ratings (at least when it is detectable in the environment), rather than as the result of a second-order noise monitoring computation, as a link to metacognition would imply. We therefore suggest that JoAs best reflect first-order comparator signals, with the metacognitive level of agency processing being second-order confidence judgments about one's agency. In addition to clarifying assumptions regarding the relationship between JoAs and metacognition, this work provides a classical confidence paradigm with which to study metacognition of agency, as well as computational models that can be used to further elucidate the mechanisms underlying explicit agency judgments.

Materials and Methods

The experiment was pre-registered (osf.io/pyjhm), and we describe deviations from the pre-registered plan.

Participants

We pre-registered a sample size of 32 participants (based on power estimates from similar tasks). We collected data until we reached 40 participants that displayed the basic and expected manipulation effect of illumination (see above). We tested 47 young, healthy participants between 18 and 35 years of age (M = 27.15, SD = 4.68) in Berlin. To participate in the study, we required that participants were right handed (Edinburgh Handedness Inventory score: M = 79.5, SD = 23.6), had no injury or condition preventing or restricting movement of the right index

finger, had normal or corrected-to-normal vision, and were fluent in English. Subjects were compensated with 8 euros per hour or with course credit and gave signed, informed consent before starting the experiment. The local ethics board approved the study (Nr. 2020-29), which conformed to the Declaration of Helsinki.

Setup

We used a LEAP Motion controller (Leap Motion Inc., San Francisco, CA) to track participants' hand motion and to control in real time the movement of a virtual hand displayed on the screen. The experiment ran on a Dell Latitude 5591 laptop (Intel core i5 with 16GB of RAM) with a display resolution of 1,920 x 1,080 (refresh rate = 60Hz) using software built in Unity 5.6.1, and was modified from software used in previous studies (Krugwasser et al., 2019; Stern et al., 2020). The computer was placed to the left of an opaque board, which occluded participants' right hand from view. Participants placed their right hand under the LEAP Motion tracker, which was fixed with its sensors facing downward. Blackout curtains were used during all testing to keep the lighting conditions within the room as consistent as possible.

Procedure

The tasks we used built on a paradigm in which participants see on the screen a virtual hand that follows the movement of their own, with a given temporal lag (Krugwasser et al., 2019). This paradigm allowed us to examine a situation closer to the more embodied or 'narrow' sense of agency that relates to control of the body itself (M.S. Christensen & Grünbaum, 2018; Dogge et al., 2019; Stern et al., 2020). Each participant completed two tasks: a confidence-rating task (Fig. 1B) and an agency-rating task (Fig. 2A). In both tasks, we manipulated sensory uncertainty by controlling the visibility of the virtual hand, allowing us to compare the effect of noise on JoAs and confidence. Importantly, we manipulated sensory noise orthogonally to the decision variable, as done by previous work (Bang & Fleming, 2018; Gardelle & Mamassian, 2015; Spence et al., 2016), which allowed us to precisely examine effects of noise without altering the true degree of control that the participants had over the virtual hand movement.

Prior to starting the experiment, participants completed the Edinburgh handedness scale. We then did a short thresholding procedure to set the illumination level that would be used in the high-noise condition of the main tasks, in order to account for differences in eyesight and

lighting conditions in the room. Participants placed their right hand on the table, under the LEAP tracker, and held it still. On each trial, participants first saw a fixation cross, followed by two consecutive presentations (separated by a flashed grey screen) of the virtual hand on the screen in the dark illumination condition, and in one case it was artificially enlarged. Participants then discriminated which of the two intervals contained the larger hand. We ran this in blocks of 10 trials and adjusted the brightness setting of the screen after each block until participants achieved 70-80 % correct in a block, and additionally did not report discomfort from straining to see the hand. This thresholding procedure took approximately 5 minutes. The brightness was, however, further adjusted if participants reported not being able to see the virtual hand movement during the training or at the beginning of the task. The brightness was only re-adjusted prior to the confidence task for one participant.

Agency-rating Task. All participants performed the agency-rating task first so that subjective ratings would not be biased by the structure or ratings of the confidence task. On each trial, participants began with their right hand resting on the table, palm facing up, with fingers extended. They saw a fixation cross (for 1.5 s), then the virtual hand appeared and they had 2 seconds in which to flex and extend their index finger once. The virtual hand displayed their movement either in real time, or with an added delay of either 70, 100, or 200 ms. Participants then rated their agency over the virtual hand movement on a scale from 1 (lowest) to 6 (highest). We explained that the term 'agency' referred to how much control they had over the movement of the virtual hand, and they were asked to focus specifically on the timing of the movement. Agency ratings were made using arrow keys to move a cursor, which started at a random position on the six-point scale each trial. Additionally, error trials in which there was a glitch of the virtual hand (such as flipping or contorting), participants saw no virtual hand, or participants made the wrong hand movement, could be marked using the Space key. Overall, 2.4% of trials were marked as errors in the confidence task, and 1.9% were marked as errors in the agency task.

To achieve the high-noise condition, the virtual hand was displayed under dark, low contrast illumination, using a directional light intensity of 0 in the Unity environment. In the low-noise condition, the virtual hand was displayed under brighter, higher contrast illumination using a directional light intensity of 0.001 (Fig. 1A). The noise conditions as well as the four delay levels were counterbalanced and randomly distributed across each block. There were 60 trials per delay level and lighting condition, for a total of 480 trials, split across 6 blocks. Prior to

this task, participants completed a short training consisting of 20 trials and including both noise conditions and all delays, but they never received any feedback regarding any task. The agency-rating task took approximately 45 minutes.

Confidence-rating task. After the agency-rating task, participants did a confidence-rating task. Participants again flexed and extended their right index finger under the LEAP motion tracker, while looking at the virtual hand on the screen. In contrast to the agency-rating task, they made two consecutive movements, each cued by the appearance of the virtual hand, and separated with a blank grey screen. They then decided which virtual hand movement they had more agency over, and rated their confidence in their response from 1 (lowest) to 6 (highest). The difference in delay levels between the movements in each trial was staircased, with one of the two movements always having no delay and the other being adjusted according to an online 2-down-1-up staircasing procedure aiming to achieve an overall accuracy of approximately 71%. Only the low-noise condition was staircased, and the delays of the high-noise condition were set to match those of the low-noise. Participants made their decision and then confidence rating using the arrow keys, and an error trial could be marked during either the decision or confidence rating.

We manipulated noise in the same way as in the agency-rating task, and this was fully counterbalanced with which movement was the delayed one, with these factors randomly distributed across each block. There were 100 trials per noise condition, for a total of 200 trials across 5 blocks. Prior to this task, participants completed another short training consisting of 10 trials, but only in the low-noise condition, to adjust to the new movement and response structure. The confidence task took approximately 45 minutes. At the end of the session, participants completed an informal debriefing.

Analysis

We removed any trials marked as errors, and any trials with reaction times shorter than 100 ms or longer than 8 s for any decision or rating.

We tested our main hypotheses using the ‘lme4’ package (Bates et al., 2015) in R (R Core Team, 2020) to build linear mixed-effects models. All models included by-participant random intercepts, and the model for the agency-rating task included random effects for the interaction of interest (Table 1). All hypotheses were tested using two-tailed tests and an alpha level of 0.05,

and additionally using Bayes factors, which we computed with the 'BayesTestR' package (Makowski et al., 2019) using default priors. To compute Bayes factors for the logistic mixed-effects analyses, we built Bayesian models with the 'brms' package (Bürkner, 2017). For each of these Bayesian regressions, we ran 4 chains of 15,000 iterations, including 5,000 burn-in samples for a total of 40,000 effective samples, and ensuring a R-hat close to 1. Effect sizes for results of the linear mixed-effects analyses were computed as η^2_p using the 'effectsize' package (Ben-Shachar et al., 2020), with 95% confidence intervals reported when possible (large sample sizes resulted in some confidence intervals of width zero, and hence uninterpretable). The results of our linear mixed-effects analyses on confidence and JoAs were confirmed using ordinal models, built using the 'ordinal' package (R.H.B. Christensen, 2019).

To analyse metacognitive ability from the confidence task, we measured metacognitive efficiency (M-Ratio) using the HMeta-d′ toolbox (Fleming, 2017). In this analysis, for the MCMC (Markov chain Monte Carlo method) procedure we used three chains of 15,000 iterations with an additional 5000 for adaptation, thinning of 3, and default initial values from JAGS (Just Another Gibbs Sampler). We also ensured that R-hat was approximately 1 for all sampling.

As a deviation from the pre-registered analyses, we included our second test-criterion and computational modeling analyses, and excluded instead some planned analysis of the variability of ratings, as, in hindsight, we reasoned that this would not help to clarify the link between agency and metacognition. To perform the Bayesian model selection and get the PEPs, we used the 'bmsR' package (Lisi, 2021), computing model evidence using Akaike weights from the MLE analysis and using 10^6 samples.

Table 1. Syntax for the Linear Mixed-Effects Models used

Task	Hypothesis	Model Formula
Confidence Task	Sensory noise influences response accuracy	logit(Response Accuracy) ~ Noise + (1 | Participant)
Confidence Task	Sensory noise influences confidence following correct decisions	Confidence ~ Response + Response Accuracy*Noise + (1 | Participant)
Agency Rating Task	Sensory noise influences the effect of delay on JoA	JoA ~ Delay*Noise + (Delay:Noise | Participant)

Modeling

To test whether JoAs reflect metacognitive computations, we compared two computational models which could both account for the observed effect of noise on agency ratings. Both are based on signal detection theory and a comparator model of agency, with the amount of delay between the real movement and virtual hand movement as the signal. However, under one model (the Bayesian-agency model) agency ratings involve a second-order assessment of sensory noise in the same way that confidence judgments do; whereas under the second one (the Rescaling model), agency is based on only first-order estimates of the internal signal strength, with the effect of noise captured by participants rescaling their ratings per condition without making metacognitive estimates of the noise.

Under both models, JoAs result from a Yes/No decision of whether participants felt agency over the virtual hand movement, and this is scaled into a rating according to a function of the strength of the evidence. Modeling JoAs as involving this binary detection decision allowed us to examine whether agency ratings follow the computations involved in decision confidence, and is also in line with work treating agency as a binary judgment (Fukushima et al., 2013; Spaniel et al., 2016). We assume that participants set an internal decision criterion (c) which determines whether they detected a delay — thus judging a disruption in their agency —, or whether they detected no delay and therefore judged themselves to have agency over the virtual hand movement. Then, the different agency ratings are modeled as additional criteria on either side of (c). We model agency ratings as getting more extreme as the perceived signal (x) gets further from the decision criterion in either direction, or in other words, as evidence supporting the agency decision increases. This predicts that perceived delays that are very long relative to

the decision criterion would lead to a 'No' decision with strong evidence, and in turn low JoAs, whereas perceived delays that are very short would lead to a 'Yes' decision with strong evidence, and high JoAs. Crucially, the two models differed in the function of internal evidence ($f(x)$) that determined the agency ratings, and in particular in the way this function was affected by sensory noise.

Agency Ratings in the Bayesian Model. The Bayesian-agency model assumed that agency ratings scale as a function of internal evidence in the same way as confidence, namely scaling with the posterior probability of being correct, given a choice and the internal signal (Sanders et al., 2016). Therefore, in this model, the agency rating is computed by estimating the probability that the agency detection was correct, given the perceived signal and detection decision. Because this probability computation depends on the level of sensory noise, the Bayesian-agency model predicts that noise will be factored into participants' JoAs.

In both models, we obtained the criterion values that split the continuous range of possible $f(x)$ values into equidistant bins. For the Bayesian model, because confidence reflects a probability, it is naturally bounded to 1. So for the Bayesian model (with a 3:3 mapping, see below) this amounted to finding the criterion values that would lead to confidence levels of ⅓, ⅔, and 1. To estimate the positions of the criteria on the internal signal axis, we followed an analytical solution that defines confidence as

$$\begin{cases} \Phi\left(\frac{c - x}{\sigma}\right) & \text{if } x \geq c \\ 1 - \Phi\left(\frac{c - x}{\sigma}\right) & \text{if } x < c \end{cases} \quad (1)$$

which we implemented, for convenience, as in a previous study (Navajas et al., 2017):

$$\Phi\left(\frac{|x - c|}{\sigma}\right) \quad (2)$$

This confidence measure can be interpreted as the perceived probability that the true delay signal was on the same side of the decision criterion as the internal signal, hence making the decision correct (Fig. 6).

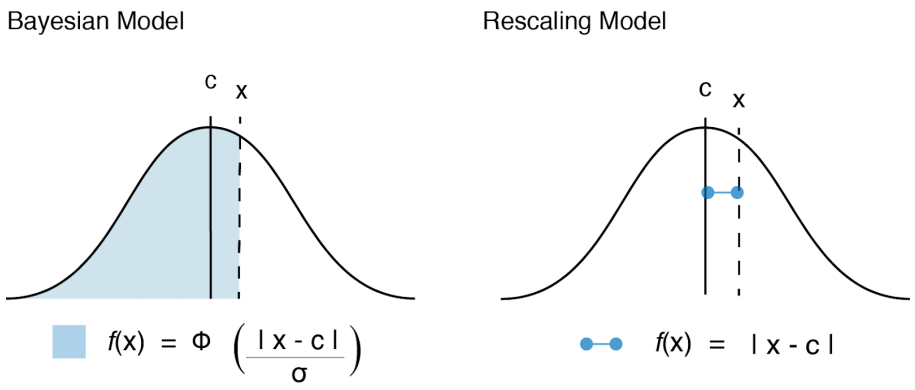


Figure 6: Function of internal evidence as estimated by each of the models tested. Agency as a function of evidence strength computations in each model. In the Bayesian model this function of internal evidence reflects confidence based on the posterior probability of having given a correct response, given internal signal and choice. In the Rescaling model this is based on the perceived distance between signal and criterion.

Agency Ratings in the Rescaling Model. The non-Bayesian alternative model considers participants to compute their ratings proportionally to the distance between a point estimate of the internal signal and the decision criterion (Fig. 6). According to this model, participants do not use the full distribution of internal signals in their assessment and hence do not make any metacognitive assessment of the precision of their evidence, but rather provide a rating that varies linearly as a function of the internal evidence, according to

$$f(x) = |x - c| \quad (3)$$

Unlike the Bayesian model, ratings in this linear model are not inherently bounded at 1, as $f(x)$ could be as high as any arbitrarily high internal signal (x). Therefore, in order to find criterion placements that divide the continuous range of $f(x)$ values into equidistant bins, we needed to approximate a maximum. We assumed that participants bound their ratings based on the range of delays they experience throughout the experiment. Because we cannot know the true range of perceived delays, we approximated the most extreme perceived delays as the most extreme external signals plus a multiple of the noise, the freely fit *scale range* parameter (s). Based on the idea that participants rescaled their agency criteria according to the noise condition that they observed, accounting for the observed behavioural interaction effect, (s) acted as a multiple of the noise within a given condition. Hence, the scale range on low noise trials would be from $-s\sigma_L$ to $[200 + s\sigma_L]$ but the scale range on high noise trials would be from $-s\sigma_H$ to $[200 + s\sigma_H]$. Fitting this parameter allowed us to obtain the maximum $f(x)$ value as the maximum distance between

(c) and either bound of the scale, and then divide the continuous *f*(x) values into equidistant bins just as we did with the Bayesian-agency model. However, due to the different *f*(x) computations, in this case the bins were equal linear distances on the internal signal axis, not equal probability bins as in the Bayesian model (Fig. 3A). Importantly, although the agency ratings do depend on the noise level due to the rescaling in this model, this does not involve participants making an assessment of the precision of their evidence, but just reflects participants considering each trial relative to a maximum that is different between illumination conditions. In other words, it would require less extreme evidence to lead to a '1' in the low-noise condition than a '1' in the high-noise condition.

Once we found criterion locations for each noise condition for each model, we calculated the probability of each rating for any given alteration and noise level. Using these probabilities, for all trials of a given participant or the pooled data, we built the likelihood function as

$$\prod_{\alpha=1}^{6} \left(\Phi\left(\frac{\gamma_{\alpha+1} - d}{\sigma}\right) - \Phi\left(\frac{\gamma_{\alpha} - d}{\sigma}\right)\right)^{n_\alpha} \qquad (4)$$

where α indexes the agency rating criterion in a given noise condition; γ_α is the position of criterion α, with γ_1 being -∞ and γ_7 being +∞; d is the external delay; and n_α is the number of trials observed for that rating and delay, in that noise condition. We then took the product of this likelihood across all four possible external delays and across both noise conditions.

Model Parameters. Both models shared the parameters: standard deviation of the low noise condition, σ_L, standard deviation of the high noise condition, σ_H, and decision criterion (c). The Rescaling model also included the scale range parameter (s). Additionally, instead of assuming that participants always used half of the scale ratings to reflect detection of agency (JoA = 4:6), and half to reflect disruption to agency (JoA = 1:3), we fit a mapping parameter to capture the number of ratings used for each decision. We fit this parameter, NRYes, defined as the number of ratings used for 'Yes' decisions, with a minimum of one rating used for each decision. If NRYes is two, for example, this would suggest participants used ratings of '5' and '6' to indicate detections of agency, and ratings of '1' to '4' to indicate disturbances to their agency. By fitting this parameter, we avoided having to make any strong assumptions about how participants used the rating scale, considering we did not have their true Yes/No decisions.

Modeling Confidence. We also applied these two models to confidence ratings, in order to compare confidence computations with those underlying JoAs. For this analysis, the models were kept the same, except instead of fitting agency criteria, we fit the confidence criteria that divided confidence ratings into 12 total bins, with 6 ratings on each side of the decision criterion. We did not need to fit NRYes, as the assignment of ratings to a particular decision was forced by the task.

Acknowledgments

We thank Angeliki Charalampaki for discussions on the work presented here, and Matthias Guggenmos and Nathan Faivre for comments on an earlier version of this manuscript. MC was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - 337619223 / RTG2386. MC and EF were supported by a Freigeist Fellowship to EF from the Volkswagen Foundation (grant number 91620). RS was supported by an Israeli Science Foundation Grant (ISF 1169/17). The funders had no role in the conceptualization, design, data collection, analysis, decision to publish, or preparation of the manuscript.

Competing Interests

The authors declare no competing interests.

Data Availability

Raw data is publicly available under https://gitlab.com/MarikaConstant/metaAgency.

Code Availability

Reproducible analysis scripts and models are publicly available under https://gitlab.com/MarikaConstant/metaAgency.

Author Contributions

M.C. and E.F. conceived and designed the study. M.C. collected and analysed the data. M.C. and R.S. contributed materials and tools. All authors interpreted the results. M.C. and E.F. wrote the paper with feedback from R.S.

References

Adler, W. T., & Ma, W. J. (2018). Limitations of Proposed Signatures of Bayesian Confidence. *Neural Computation*, *30*(12), 3327–3354. https://doi.org/10.1162/neco_a_01141

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*, No Pagination Specified-No Pagination Specified. https://doi.org/10.1037/xge0000892

Bang, D., & Fleming, S. M. (2018). Distinct encoding of decision confidence in human medial prefrontal cortex. *Proceedings of the National Academy of Sciences*, *115*(23), 6082–6087. https://doi.org/10.1073/pnas.1800795115

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Ben-Shachar, M., Lüdecke, D., & Makowski, D. (2020). effectsize: Estimation of Effect Size Indices and Standardized Parameters. *Journal of Open Source Software*, *5*(56), 2815. https://doi.org/10.21105/joss.02815

Bürkner, P.-C. (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Carruthers, G. (2012). The case for the comparator model as an explanation of the sense of agency and its breakdowns. *Consciousness and Cognition*, *21*, 30–45; discussion 55. https://doi.org/10.1016/j.concog.2010.08.005

Christensen, M. S., & Grünbaum, T. (2018). Sense of agency for movements. *Consciousness and Cognition*, *65*, 27–47. https://doi.org/10.1016/j.concog.2018.07.002

Christensen, R. H. B. (2019). *ordinal: Regression Models for Ordinal Data* (2019.12-10) [Computer software]. https://CRAN.R-project.org/package=ordinal

Corlett, P. R., Taylor, J. R., Wang, X.-J., Fletcher, P. C., & Krystal, J. H. (2010). Toward a neurobiology of delusions. *Progress in Neurobiology*, *92*(3), 345–369.

https://doi.org/10.1016/j.pneurobio.2010.06.007

Dewey, J. A., & Knoblich, G. (2014). Do Implicit and Explicit Measures of the Sense of Agency Measure the Same Thing? *PLOS ONE*, *9*(10), e110118. https://doi.org/10.1371/journal.pone.0110118

Dogge, M., Custers, R., & Aarts, H. (2019). Moving Forward: On the Limits of Motor-Based Forward Models. *Trends in Cognitive Sciences*, *23*(9), 743–753. https://doi.org/10.1016/j.tics.2019.06.008

Flavell, J. H. (1978). Metacognitive development. In J. M. Scandura & C. J. Brainerd (Eds.), *Structural/process theories of complex human behaviour* (pp. 213–247). Sijthoff & Noordhoff.

Fleming, S. M. (2017). HMeta-d: Hierarchical Bayesian estimation of metacognitive efficiency from confidence ratings. *Neuroscience of Consciousness*, *2017*(nix007). https://doi.org/10.1093/nc/nix007

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. https://doi.org/10.1098/rstb.2012.0021

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

Fletcher, P. C., & Frith, C. D. (2009). Perceiving is believing: A Bayesian approach to explaining the positive symptoms of schizophrenia. *Nature Reviews Neuroscience*, *10*(1), 48–58. https://doi.org/10.1038/nrn2536

Frith, C. D., Blakemore, S. J., & Wolpert, D. M. (2000). Abnormalities in the awareness and control of action. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *355*(1404), 1771–1788.

Fukushima, H., Goto, Y., Maeda, T., Kato, M., & Umeda, S. (2013). Neural Substrates for Judgment of Self-Agency in Ambiguous Situations. *PLOS ONE*, *8*(8), e72267. https://doi.org/10.1371/journal.pone.0072267

Gallagher, S. (2007). The Natural Philosophy of Agency. *Philosophy Compass*, *2*(2), 347–357. https://doi.org/10.1111/j.1747-9991.2007.00067.x

Gardelle, V. de, & Mamassian, P. (2015). Weighting Mean and Variability during Confidence Judgments. *PLOS ONE*, *10*(3), e0120870. https://doi.org/10.1371/journal.pone.0120870

Haggard, P. (2017). Sense of agency in the human brain. *Nature Reviews Neuroscience*, *18*(4), 196–207. https://doi.org/10.1038/nrn.2017.14

Haggard, P., & Tsakiris, M. (2009). The Experience of Agency: Feelings, Judgments, and Responsibility. *Current Directions in Psychological Science*, *18*(4), 242–246. https://doi.org/10.1111/j.1467-8721.2009.01644.x

Imaizumi, S., & Tanno, Y. (2019). Intentional binding coincides with explicit sense of agency. *Consciousness and Cognition*, *67*, 1–15. https://doi.org/10.1016/j.concog.2018.11.005

Jamieson, D. G., & Petrusic, W. M. (1975). Presentation order effects in duration discrimination. *Perception & Psychophysics*, *17*(2), 197–202. https://doi.org/10.3758/BF03203886

Kawabe, T., Roseboom, W., & Nishida, S. (2013). The sense of agency is action–effect causality perception based on cross-modal grouping. *Proceedings of the Royal Society B: Biological Sciences*, *280*(1763), 20130991. https://doi.org/10.1098/rspb.2013.0991

Kepecs, A., & Mainen, Z. F. (2012). A computational framework for the study of confidence in humans and animals. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1322–1337. https://doi.org/10.1098/rstb.2012.0037

Krugwasser, A. R., Harel, E. V., & Salomon, R. (2019). The boundaries of the self: The sense of agency across different sensorimotor aspects. *Journal of Vision*, *19*(4), 14–14. https://doi.org/10.1167/19.4.14

Legaspi, R., & Toyoizumi, T. (2019). A Bayesian psychophysics model of sense of agency. *Nature Communications*, *10*(1), 4250. https://doi.org/10.1038/s41467-019-12170-0

Macmillan, N. A., & Creelman, C. D. (1991). *Detection theory: A user's guide* (pp. xv, 407). Cambridge University Press.

Makowski, D., Ben-Shachar, M. S., & Lüdecke, D. (2019). bayestestR: Describing Effects and their Uncertainty, Existence and Significance within the Bayesian Framework. *Journal of Open Source*

Software, *4*(40), 1541. https://doi.org/10.21105/joss.01541

Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition*, *21*(1), 422–430. https://doi.org/10.1016/j.concog.2011.09.021

Lisi, M. (2021). bmsR: Bayesian model selection for group studies in R. R package version 0.0.1.0000.

Metcalfe, J., & Greene, M. J. (2007). Metacognition of agency. *Journal of Experimental Psychology. General*, *136*(2), 184–199. https://doi.org/10.1037/0096-3445.136.2.184

Metcalfe, J., Van Snellenberg, J. X., DeRosse, P., Balsam, P., & Malhotra, A. K. (2012). Judgements of agency in schizophrenia: An impairment in autonoetic metacognition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1391–1400. https://doi.org/10.1098/rstb.2012.0006

Meyniel, F., Sigman, M., & Mainen, Z. F. (2015). Confidence as Bayesian Probability: From Neural Origins to Behavior. *Neuron*, *88*(1), 78–92. https://doi.org/10.1016/j.neuron.2015.09.039

Miele, D. B., Wager, T. D., Mitchell, J. P., & Metcalfe, J. (2011). Dissociating Neural Correlates of Action Monitoring and Metacognition of Agency. *Journal of Cognitive Neuroscience*, *23*(11), 3620–3636. https://doi.org/10.1162/jocn_a_00052

Minohara, R., Wen, W., Hamasaki, S., Maeda, T., Kato, M., Yamakawa, H., Yamashita, A., & Asama, H. (2016). Strength of Intentional Effort Enhances the Sense of Agency. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.01165

Moore, J. W., & Fletcher, P. C. (2012). Sense of agency in health and disease: A review of cue integration approaches. *Consciousness and Cognition*, *21*(1), 59–68. https://doi.org/10.1016/j.concog.2011.08.010

Moore, J. W., & Haggard, P. (2008). Awareness of action: Inference and prediction. *Consciousness and Cognition*, *17*(1), 136–144. https://doi.org/10.1016/j.concog.2006.12.004

Moore, J. W., Wegner, D. M., & Haggard, P. (2009). Modulating the sense of agency with external cues. *Consciousness and Cognition*, *18*(4), 1056–1064. https://doi.org/10.1016/j.concog.2009.05.004

Navajas, J., Hindocha, C., Foda, H., Keramati, M., Latham, P. E., & Bahrami, B. (2017). The idiosyncratic nature of confidence. *Nature Human Behaviour*, *1*(11), 810–818. https://doi.org/10.1038/s41562-017-0215-1

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

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

Raftery, A. E. (1995). Bayesian Model Selection in Social Research. *Sociological Methodology*, *25*, 111–163. https://doi.org/10.2307/271063

Rausch, M., Hellmann, S., & Zehetleitner, M. (2018). Confidence in masked orientation judgments is informed by both evidence and visibility. *Attention, Perception & Psychophysics*, *80*(1), 134–154. https://doi.org/10.3758/s13414-017-1431-5

Rausch, M., & Zehetleitner, M. (2019). The folded X-pattern is not necessarily a statistical signature of decision confidence. *PLOS Computational Biology*, *15*, e1007456. https://doi.org/10.1371/journal.pcbi.1007456

Rigoux, L., Stephan, K. E., Friston, K. J., & Daunizeau, J. (2014). Bayesian model selection for group studies—Revisited. NeuroImage, 84, 971–985. https://doi.org/10.1016/j.neuroimage.2013.08.065

Robinson, J. D., Wagner, N.-F., & Northoff, G. (2016). Is the Sense of Agency in Schizophrenia Influenced by Resting-State Variation in Self-Referential Regions of the Brain? *Schizophrenia Bulletin*, *42*(2), 270–276. https://doi.org/10.1093/schbul/sbv102

Sanders, J. I., Hangya, B., & Kepecs, A. (2016). Signatures of a Statistical Computation in the Human Sense of Confidence. *Neuron*, *90*(3), 499–506. https://doi.org/10.1016/j.neuron.2016.03.025

Sato, A. (2009). Both motor prediction and conceptual congruency between preview and action-effect contribute to explicit judgment of agency. *Cognition*, *110*(1), 74–83. https://doi.org/10.1016/j.cognition.2008.10.011

Sato, A., & Yasuda, A. (2005). Illusion of sense of self-agency: Discrepancy between the predicted and actual sensory consequences of actions modulates the sense of self-agency, but not the sense of self-ownership. *Cognition*, *94*(3), 241–255. https://doi.org/10.1016/j.cognition.2004.04.003

Spaniel, F., Tintera, J., Rydlo, J., Ibrahim, I., Kasparek, T., Horacek, J., Zaytseva, Y., Matejka, M., Fialova, M., Slovakova, A., Mikolas, P., Melicher, T., Görnerova, N., Höschl, C., & Hajek, T. (2016). Altered Neural Correlate of the Self-Agency Experience in First-Episode Schizophrenia-Spectrum Patients: An fMRI Study. *Schizophrenia Bulletin*, *42*(4), 916–925. https://doi.org/10.1093/schbul/sbv188

Spence, M. L., Dux, P. E., & Arnold, D. H. (2016). Computations underlying confidence in visual perception. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(5), 671–682. https://doi.org/10.1037/xhp0000179

Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J., & Friston, K. J. (2009). Bayesian Model Selection for Group Studies. NeuroImage, 46(4), 1004–1017. https://doi.org/10.1016/j.neuroimage.2009.03.025

Stern, Y., Koren, D., Moebus, R., Panishev, G., & Salomon, R. (2020). Assessing the Relationship between Sense of Agency, the Bodily-Self and Stress: Four Virtual-Reality Experiments in Healthy Individuals. *Journal of Clinical Medicine*, *9*(9). https://doi.org/10.3390/jcm9092931

Synofzik, M., Vosgerau, G., & Newen, A. (2008). Beyond the comparator model: A multifactorial two-step account of agency. *Consciousness and Cognition*, *17*(1), 219–239. https://doi.org/10.1016/j.concog.2007.03.010

Synofzik, M., Vosgerau, G., & Voss, M. (2013). The experience of agency: An interplay between prediction and postdiction. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00127

Voss, M., Chambon, V., Wenke, D., Kühn, S., & Haggard, P. (2017). In and out of control: Brain mechanisms linking fluency of action selection to self-agency in patients with schizophrenia. *Brain*, *140*(8), 2226–2239. https://doi.org/10.1093/brain/awx136

Wagenmakers, E.-J., & Farrell, S. (2004). AIC model selection using Akaike weights. Psychonomic

Bulletin & Review, 11(1), 192–196. https://doi.org/10.3758/BF03206482

Wang, S., Rajananda, S., Lau, H., & Knotts, J. D. (2020). New measures of agency from an adaptive sensorimotor task. *PloS One*, *15*(12), e0244113. https://doi.org/10.1371/journal.pone.0244113

Wen, W. (2019). Does delay in feedback diminish sense of agency? A review. *Consciousness and Cognition*, *73*, 102759. https://doi.org/10.1016/j.concog.2019.05.007

Wen, W., Yamashita, A., & Asama, H. (2015). The influence of action-outcome delay and arousal on sense of agency and the intentional binding effect. *Consciousness and Cognition*, *36*, 87–95. https://doi.org/10.1016/j.concog.2015.06.004

Wenke, D., Fleming, S. M., & Haggard, P. (2010). Subliminal priming of actions influences sense of control over effects of action. *Cognition*, *115*(1), 26–38. https://doi.org/10.1016/j.cognition.2009.10.016

Wolpe, N., Haggard, P., Siebner, H. R., & Rowe, J. B. (2013). Cue integration and the perception of action in intentional binding. *Experimental Brain Research*, *229*(3), 467–474. https://doi.org/10.1007/s00221-013-3419-2

Yeshurun, Y., Carrasco, M., & Maloney, L. T. (2008). Bias and sensitivity in two-interval forced choice procedures: Tests of the difference model. *Vision Research*, *48*(17), 1837–1851. https://doi.org/10.1016/j.visres.2008.05.008

Supplementary Information

Confidence Task Regression Analysis - Ordinal Models

Because participants rated confidence on a discrete scale from 1-6, we reran the linear mixed effects (LME) analysis from the Confidence Task using ordinal models. We built a model on confidence ratings including the interaction between response accuracy and noise level and each factor as fixed effects, response identity as an additional fixed effect, as well as by-participant random intercepts. This confirmed the results of our analysis using linear regression, and revealed a significant interaction between Response Accuracy and Noise, $\chi^2(1) = 14.86$, $p < 0.001$. This also confirmed the significant main effect of Response, $\chi^2(1) = 78.91$, $p < 0.001$.

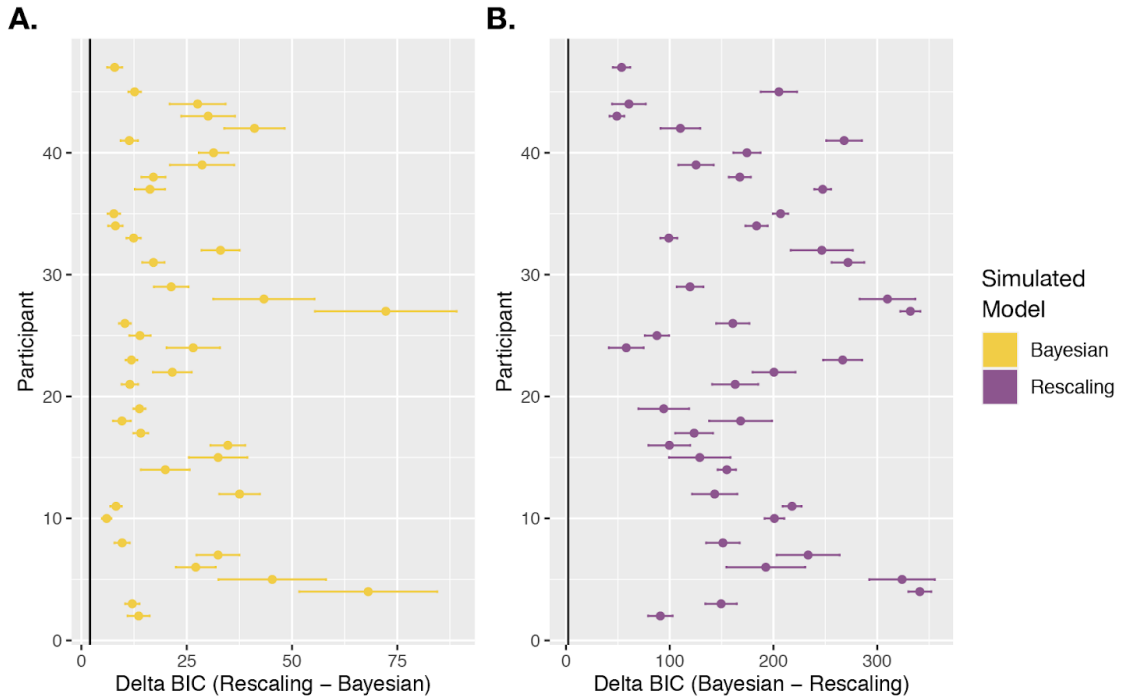
Agency Rating Task Regression Analysis - Ordinal Models

We also reran the LME analysis from the Agency Rating Task with ordinal models. We built a model on JoAs including the noise level and delay, and their interaction as fixed effects, as well as by-participant random effects of the interaction, and random intercepts. This confirmed the results of our LME analysis, revealing a significant interaction between Noise and Delay on JoAs, $\chi^2(1) = 43.15$, $p < 0.001$, as well as a significant main effect of Delay $\chi^2(1) = 23.56$, $p < 0.001$, replicating previous work (Krugwasser et al., 2019; Stern et al., 2020).

Model Recovery Analysis

To confirm that the Bayesian and Rescaling models were distinguishable from one another, we ran a model recovery analysis. We took the winning parameters for each model for each participant and simulated 20 datasets with each parameter set. This generated 800 simulated datasets per model, covering all of the relevant parameters of our real participants. We then fit both models to those 1600 simulated datasets and performed BIC comparisons as we did in our real data. We required a minimum BIC difference of 2 for one model to be considered a better fit, just as we did in our main modeling analyses. This revealed the Bayesian model to be the better fitting model for 776 out of 800 of the datasets simulated with the Bayesian model, suggesting it to be correctly recovered 97.00% of the time overall. Similarly, the Rescaling model was the better fitting model for 799 out of 800 of the datasets simulated with the Rescaling model, being

correctly recovered 99.88% of the time overall. The mean BIC difference from models fit to 20 repetitions of each participant's relevant parameter set for each model are shown in the figure below. The mean ΔBIC (from the 20 repetitions) was above the minimum cutoff of 2, favoring the correct model, for all participants' parameter settings, for both models. Together, this analysis confirms the models to be distinguishable from one another across the entire relevant parameter space.



Model Recovery Results. BIC differences between the Bayeseian and Rescaling models, with positive values favoring the simulated model, fit to simulated datasets using all participants' winning parameter sets for each model. **(A.)** BIC comparison results from datasets simulated with the Bayesian model and the winning parameters for the Bayesian model for each participant. **(B.)** BIC comparison results from datasets simulated with the Rescaling model and the winning parameters for the Rescaling model for each participant. Points reflect the mean BIC difference across 20 repetitions of each parameter set, and error bars show the standard error. The black vertical line indicates the minimum BIC difference of 2 required for the correct model to be considered the better fitting model.

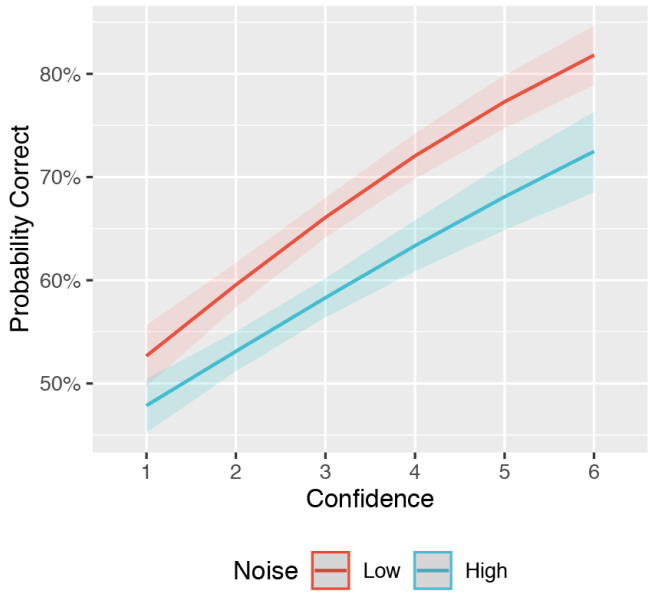
Measuring Metacognitive Sensitivity with Logistic Regressions

Along with our M-Ratio analysis, we also measured metacognitive sensitivity using logistic regressions, which does not account for first-order task performance. We built a mixed logistic regression model on Response Accuracy including Confidence and Noise as well as their interaction as fixed effects, and random Confidence slopes and random intercepts per participant

(see below for model syntax). Metacognitive performance was estimated by considering the effect of confidence as a measure of how well confidence ratings tracked accuracy. We found a significant main effect of Confidence ($\chi^2(1) = 45.75$, $p < 0.001$, $BF_{10} = 1.24 \times 10^9$), which confirmed that participants' confidence ratings tracked their accuracy and were hence meaningful, despite the difficult task. Further, in line with the manipulation check displayed in Figure 1C (in the main text), we found a main effect of Noise on accuracy ($\chi^2(1) = 74.74$, $p < 0.001$, $BF_{10} = 4.14 \times 10^{12}$). The data are inconclusive on whether the two factors interact, as a frequentist analysis revealed a significant interaction between Confidence and Noise ($\chi^2(1) = 5.22$, $p = 0.022$), whereas Bayesian statistics revealed no conclusive evidence either way ($BF_{10} = 1.01$).

Model Syntax

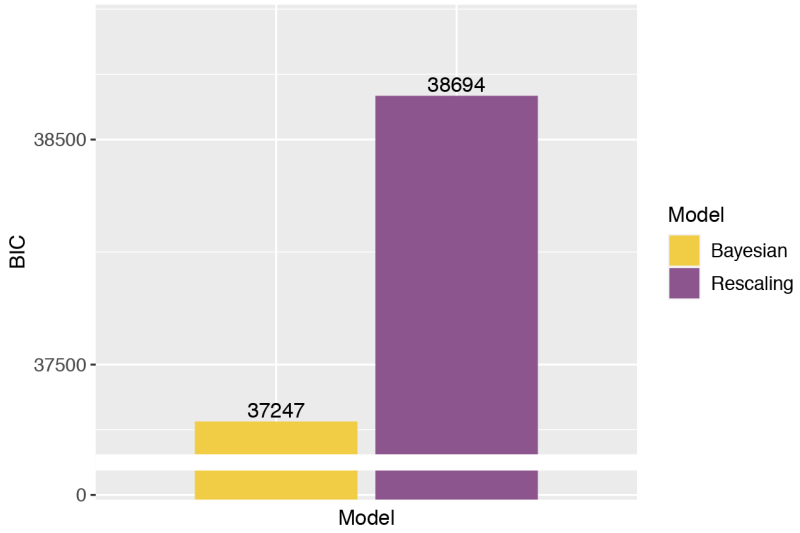
Task	Hypothesis	Model Formula
Confidence Task	Confidence tracks response accuracy, modulated by sensory noise	logit(Response Accuracy) ~ Confidence*Noise + (Confidence | Participant)



Logistic Regression Results. Metacognitive sensitivity quantified with predicted probability correct across confidence ratings and noise conditions from mixed logistic regression model results. We found a significant interaction between noise and confidence, with a smaller slope of confidence across accuracy under high noise. 95% confidence intervals shown.

Confidence Model-Fitting Using Additional Information from Confidence Task

The confidence task provided us with additional information that could be used to fit the models, that we could not access in the agency rating task, namely, the calculated decision criteria and the difference in noise level between conditions, with σ_H calculated as $\sigma_L * d'_H / d'_L$. Because we aimed to compare JoAs and confidence ratings in terms of their underlying computations, our main modeling analysis fit the two models to confidence ratings without using any of the additional information available, and subjecting the analysis to the same assumptions as with agency ratings. This involved freely fitting the noise levels, and assuming optimal decision criteria, as we did in the agency task. Here, we repeated the group-level analysis using the calculated decision criterion and noise difference. The calculated decision criteria for the pooled data were -0.079 in the low-noise condition and -0.13 in the high-noise condition. The high-noise level was calculated to be 2.73 times the low-noise, based on d'_L/d'_H. The winning parameters from this analysis for the Rescaling model were $\sigma_L = 0.73$, leaving σ_H as 2.00, and (s) = 1.30. The winning parameter for the Bayesian model was $\sigma_L = 0.37$, leaving σ_H as 1.00. The Bayesian model could still better explain confidence ratings ($\Delta BIC_{Rescaling\text{-}Bayes} = 1448$).



BIC Results from Confidence Task. BIC comparison between the Bayesian and Rescaling models on pooled confidence rating data, using calculated decision criteria and difference between noise conditions.