A Bayesian psychophysics model of sense of agency

Roberto Legaspi* and Taro Toyoizumi*

Laboratory for Neural Computation and Adaptation

RIKEN Center for Brain Science

RIKEN CBS-OMRON Collaboration Center

2-1 Hirosawa Wako City, Saitama 351-0198, Japan

Tel: +81-48-467-9644; Fax: +81-48-467-9670

*Corresponding authors: roberto.legaspi@riken.jp, taro.toyoizumi@riken.jp

Abstract

Despite the increasing significance of sense of agency (SoA) research, the literature lacks a formal model: what computational principles underlie SoA, the registration that oneself initiated an action that caused something to happen? We theorize SoA in the framework of optimal Bayesian cue integration with mutually involved principles, namely, reliability of action and outcome sensory signals, their consistency with the causation of the outcome by the action, and the prior belief in causation. We used our Bayesian model to explain the intentional binding effect, hailed as reliable indicator of SoA. Our model explains temporal binding in both self-intended and unintentional actions suggesting that intentionality is not strictly necessary given high confidence in the action causing the outcome. Our Bayesian model also explains that if the sensory cues are reliable, SoA can emerge even for unintended actions. Our formal model therefore posits a *precision*-dependent *causal* agency.

The number of scientific contributions being added to the theoretical literature of sense of agency (SoA) has significantly increased at least in the past two decades[1,2]. The concept has garnered considerable attention in psychology, philosophy, neuroscience and psychopathology[3]. SoA is the registration[4] that the self initiates an action in order to interact with and influence its external environment[5]. It has been posited that SoA is fundamental to the experience of volition[6-9] and to self-consciousness because of its self-other distinction[10-12], and the degradation of this experience characterizes certain psychiatric and neurological disorders[13-15]. Furthermore, SoA has recently been suggested to underpin neuroethics and law due to the role it plays in the social concept of responsibility for one's own actions[5,9,16,17].

Despite its increasing significance, the literature still lacks the computational principles that underlie SoA. We theorize SoA as the *confidence* in one's perception of the action-outcome effect and that it is consistent (e.g., spatially or temporally) with the hypothesis that the action caused the outcome. We adapted the model of Sato, Toyoizumi and Aihara[18] that was originally used to explain the *ventriloquism* effect as a Bayesian estimate of a common cause behind the consistency of the audiovisual stimuli. Formalizing SoA by this Bayesian psychophysics principle distinguishes our theory from existing works.

We compared the predictions of our model to the results of two pertinent intentional binding studies. Intentional binding, which is the perceived compression of the time interval between voluntary action and its outcome, has been reported as reliable implicit measure of SoA and has been used in a large number of studies providing valuable analyses on the temporal perception of action-outcome effects and the nature of SoA[19]. The seminal experiment of Haggard, Clark & Kalogeras[6] investigated the perceived action-outcome timing effects in three conditions: voluntary wherein the subject intentionally presses a button, involuntary wherein muscle twitches of the subject's hand are induced by a transcranial magnetic stimulation (TMS) applied to the motor cortex, and sham TMS wherein the TMS on the parietal cortex produces audible clicks but no movement (hereafter, voluntary, involuntary, and sham conditions, respectively). Haggard and colleagues computed the time interval between the perceived action timings (with the timings of either voluntary actions, muscle twitches, or audible TMS clicks as control experiment) and the perceived timings of subsequent tone stimuli. They showed that voluntary actions produced intentional binding, involuntary muscle twitches produced repulsion, i.e., prolonged opposite perception of the action-outcome intervals, and audible TMS clicks produced neither binding nor repulsion. Hence, they posit intentionality is necessary to achieve action-outcome binding.

The second pertains to the study of Wolpe, Haggard, Siebner & Rowe[20] that investigated the contribution of cue integration to intentional binding by manipulating the reliability of the consequent tone relative to a background white noise. Such manipulation resulted in three levels of tone uncertainty conditions, namely, low, intermediate and high uncertainty. Their analyses showed that when tone reliability was reduced, the perceptual shift in tone timing towards the action was increased.

Our Bayesian model reproduces the above empirical results based on a computational principle. Further, it goes beyond timing estimations by exposing the underlying Bayesian mechanisms that possibly drove the temporal binding. Our Bayesian model explains the perceived compressed time interval between the action-outcome effect is more *consistent* with the prior belief of the *causal* role of one's action in producing the immediate outcome, and thus increases the confidence in the Bayesian estimate assuming the causal case, modeled as SoA. Moreover, our model explains intentional binding as a specific class of the more general notion of causal binding. Our Bayesian model predicts that intentional binding generally happens on a per-trial basis, yielding a bimodal distribution of the perceived action-outcome interval. Lastly, the model also predicts that if the sensory input signals are perceived as reliable (precise), SoA may arise even for unintended actions, which serves as a testable theory for

future SoA experiments. Our theory therefore provides a formal model that coherently accounts for the relationship between time perception, perception of causality, and reliability of action and outcome cues.

Results

We considered the experimental setup of intentional binding where a subject presses a button (i.e., the action) and a tone (i.e., the outcome) sounds 250 ms after the button press. The true action and outcome timings are thus described by $t_A^* = 0$ ms and $t_O^* = 250$ ms, respectively, but they are unknown to the subject. The task for the subject is to accurately report her perceived timings of the button press and tone. We assume the arrival of relevant sensory input informing the timing of each of these physical events involves sensory delay d and jitter of variance σ^2 due to sensory noise. Thus, the arrival time τ_A of sensory input that signals the action timing is assumed to be generated from a Gaussian distribution, $\mathcal{N}(t_A^* + d_A, \sigma_A^2)$, with mean $t_A^* + d_A$ and variance σ_A^2. Similarly, the arrival time τ_O of sensory input that signals the outcome timing is generated from $\mathcal{N}(t_O^* + d_O, \sigma_O^2)$.

The brain often resolves such ambiguity in sensory inputs by integrating multiple sensory cues akin to the Bayesian "ideal observer"[21]. Hence, we model a Bayesian observer who estimates action timing t_A and outcome timing t_O based on the corresponding noisy sensory inputs arriving at time τ_A for the action and τ_O for the outcome. The conditional probability distributions of τ_A and τ_O that the Bayesian observer uses are modeled as Gaussian distributions

$$P(\tau_A|t_A) \propto \exp\left(-\frac{(\tau_A - t_A)^2}{2{\sigma_A}^2}\right)$$
$$P(\tau_O|t_O) \propto \exp\left(-\frac{(\tau_O - t_O)^2}{2{\sigma_O}^2}\right), \quad (1)$$

with mean t_A and t_O and variance σ_A^2 and σ_O^2 for action and outcome, respectively. Note that sensory delays d_A and d_O are not included in equation (1) for the reason we describe in the next paragraph.

Before studying the binding effect, let us consider simple baseline conditions. In one baseline condition, the action timing is reported by the subject without the presentation of an outcome tone. If no prior knowledge is available, the Bayesian observer reports the action timing that maximizes the conditional probability distribution in equation (1). Hence, the estimated action timing $\hat{t}_A = \tau_A$ is solely determined by the noisy sensory input informing the action timing. In this case, the model predicts that the distribution of $\hat{t}_A$ is $\mathcal{N}(t_A^* + d_A, \sigma_A^2)$. The mean and standard deviation of $\hat{t}_A$ in the baseline condition were experimentally reported, e.g., Haggard's results in the voluntary condition suggest $d_A = 6$ ms and $\sigma_A = 66$ ms (refer to Table 1 in Methods for all condition-based d_A and σ_A values). Importantly, we assume that the observer does not take into account sensory delay d_A in equation (1). If the Bayesian observer included its effect, it could compensate for this delay and report unbiased timing, which was not the case in the experiment. Therefore, we assume that the observer was unable to take into account the sensory delay in equation (1). In the other baseline condition, the subject passively listens to a tone and reports its timing. This case goes parallel to the above case, and the model predicts that the estimated tone timing is $\hat{t}_O = \tau_O$, which is distributed according to $\mathcal{N}(t_O^* + d_O, \sigma_O^2)$. The comparison of this model prediction to Haggard's experiment, for example, would be $d_O = 15$ ms and $\sigma_O = 72$ ms (refer to Table 1).

Next, we study the effect of binding when the subject makes an action and then listens to the outcome tone, commonly referred to as the operant condition. In this case, the Bayesian observer makes an inference not only based on the conditional probability distribution in equation (1) but also based on the prior distribution of t_A and t_O. Adapting the Bayesian model

of the v*entriloquism effect*[18], we assume the prior distribution depends on the observer's belief if the action caused the outcome, i.e., the *causal* case: $\xi = 1$, or the action and the outcome are unrelated, i.e., the *acausal* case: $\xi = 0$:

$$P(t_A, t_O \mid \xi) \propto \begin{cases} \exp\left(-\frac{(t_O - t_A - \mu_{AO})^2}{2{\sigma_{AO}}^2}\right), & (\xi = 1) \\ 1. & (\xi = 0) \end{cases} \quad (2)$$

The action causes the outcome in the causal case ($\xi = 1$) so that the outcome timing involves a typical delay μ_{AO} with respect to the action timing and a Gaussian-distributed jitter of standard deviation σ_{AO}. The outcome is caused by something other than the action in the acausal case ($\xi = 0$) so that t_A and t_O are independent. Lastly, we define $P(\xi)$ as the *prior* for each belief: $P(\xi = 1)$ for the causal case and $P(\xi = 0) = 1 - P(\xi = 1)$ for the acausal case. We hypothesize the estimation of ξ to be essential for the perception of causality and SoA (explained below).

The prior probability of equation (2) cannot be normalized unless a finite range of (t_A, t_O) is defined. Therefore, we only consider it in the range $R = \{t_A, t_O | t_A \in (t_A^* - T/2,\ t_A^* + T/2), t_O \in (t_O^* - T/2, t_O^* + T/2)\}$ and assume that it is zero outside R, where again $t_A^* = 0$ ms and $t_O^* = 250$ ms are the true action and outcome timings, unknown to the observer, and $T = 250$ ms is a large enough but finite constant that specify the interval lengths in consideration. Hence, the prior probability distribution $P(t_A, t_O \mid \xi)$ in equation (2) must be normalized within R. Our results are robust to a shift in the center of R.

Given a pair of sensory inputs at τ_A and τ_O, the Bayesian observer estimates the most probable timing for the action and the outcome and whether these observations are consistent with the causal case. According to the Bayesian estimation theorem, the maximum-a-posteriori (MAP) estimate $(\hat{t}_A, \hat{t}_O, \hat{\xi})$ of the corresponding pair of physical sensory timing (t_A, t_O) and the causal variable ξ is given by

$$\left(\hat{t}_A, \hat{t}_O, \hat{\xi}\right) = \arg\max_{t_A, t_O, \xi} P(t_A, t_O, \xi \mid \tau_A, \tau_O), \quad (3)$$

where $P(t_A, t_O, \xi \mid \tau_A, \tau_O)$ is the *posterior* probability distribution of (t_A, t_O, ξ) given the sensory inputs (τ_A, τ_O). Hence, whether the Bayesian observer estimates the action-outcome effect to be causal or not depends on the posterior-ratio comparing the causal case ($\xi = 1$) and the acausal case ($\xi = 0$), namely,

$$r \equiv \max_{t_A, t_O} P(t_A, t_O, \xi = 1 \mid \tau_A, \tau_O) \Big/ \max_{t_A, t_O} P(t_A, t_O, \xi = 0 \mid \tau_A, \tau_O) \quad (4)$$

Causality is detected if the confidence in the causal estimate is greater than that in the acausal case, that is, $r > 1$. The MAP estimate of equation (3) is then given by (see Methods for the derivation)

$$\left(\hat{t}_A, \hat{t}_O, \hat{\xi}\right) = \begin{cases} \left(\tau_A + \frac{\sigma_A^2}{\sigma_{tot}^2}(\tau_O - \tau_A - \mu_{AO}), \tau_O - \frac{\sigma_O^2}{\sigma_{tot}^2}(\tau_O - \tau_A - \mu_{AO}), 1\right), & (r > 1) \\ (\tau_A, \tau_O, 0), & (r < 1) \end{cases} \quad (5)$$

with $\sigma_{tot}^2 \equiv \sigma_A^2 + \sigma_O^2 + \sigma_{AO}^2$. This indicates, on one hand, that perceptual shift does not happen if the causality is not detected ($\hat{\xi} = 0$) – the time estimates for action and outcome simply reflect the corresponding sensory signals in this case. On the other hand, perceptual shift happens if the causality is detected ($\hat{\xi} = 1$) – the action and outcome timing attract each other in the form of binding if $\tau_O - \tau_A > \mu_{AO}$ and repel each other in the form of repulsion if $\tau_O - \tau_A < \mu_{AO}$. The magnitude of perceptual shift for the action and outcome timing is given by $\sigma_A^2/\sigma_{tot}^2$ and $\sigma_O^2/\sigma_{tot}^2$, respectively, implying that perceptual shift is greater for a more unreliable stimulus. This model predicts that the occurrence of binding, repulsion, or no perceptual shift is trial-dependent, influenced by the noisy sensory signal $\tau_O - \tau_A$ informing the action-outcome interval. We denote the probability of detecting causality (i.e., $\hat{\xi} = 1$) by

P_c (see Method for its analytical expression). P_c increases with larger $P(\xi = 1)$ and smaller σ_{AO} if $\sigma_{AO} \ll \sigma_A, \sigma_O$.

Lastly, separately from the judgement of causality described above, we also directly quantify the confidence in the causal MAP estimate

$$CCE = \max_{t_A, t_O} P(t_A, t_O, \xi = 1 | \tau_A, \tau_O) \quad (6)$$

and we postulate this quantity to be a possible indication of the pre-reflective feeling of agency (see Discussion). The analytical expression of CCE in the appendix yields the following requirements to have high CCE: (A) The timing of sensory signals must be consistent with the causation of the outcome by the action, namely, $\tau_O - \tau_A \approx \mu_{AO}$; (B) The causal prior probability $P(\xi = 1)$ must be high; (C) The sensory inputs must be precise, i.e., the amplitudes σ_A and σ_O of sensory jitter must be small enough. We therefore posit SoA as encapsulation and manifestation of several pertinent aspects, which include temporal consistency in the action-outcome effect, the prior belief of an action causing the outcome, and the reliability of the perceived sensory signals.

Here, we briefly describe how we obtained the parameter values used in our simulation (but see Methods for more details about the simulation and model fitting). Fitting of d_A,d_O, σ_A and σ_O is straightforward, they are suggested by the means and standard deviations of the reported subjects' baseline estimation errors (Table 1-Sets A and B). After fixing these parameters, the model is left with three free parameters, μ_{AO}, σ_{AO}, and $P(\xi = 1)$. As described in equation (5), μ_{AO} has an important role in determining if binding or repulsion happens in each experimental condition. A fixed value of μ_{AO} = 230 ms successfully accounts for this qualitative behavior in all the 6 experimental conditions (3 from Haggard et al. and 3 from Wolpe et al.) that we study. The analytical expressions in Methods suggest that σ_{AO} and $P(\xi = 1)$ have a largely overlapping role in detecting causality. Causality is more likely detected if σ_{AO} is small or $P(\xi = 1)$ is large, although the exact mechanisms are slightly different. At least one of these two parameters needs to be adjusted according to the conditions to account for the experimental observations. For simplicity, we fix σ_{AO} = 10 ms to be a small enough constant to permit noticeable perceptual shift and adjust $P(\xi = 1)$ (see Table 1 for the parameter values in 6 experimental conditions) to account for two observations in each condition, namely, the perceptual shifts in the action timing and the outcome timing.

Our results show that our simple Bayesian model qualitatively reproduces the perceptual shifts that were reported in Haggard et al.'s study (Fig. 1). Consistent to their findings, our Bayesian observer inferred the action and outcome perceptual shifts to bind in the voluntary condition resulting to compressed temporal intervals between the action and outcome perceptual shifts. However, repulsion, i.e., reversed and prolonged perceptual shifts, was observed for the involuntary condition. The model also reproduced no appreciable perceptual shifts in the sham condition.

Our Bayesian model predicts binding and repulsion to increase with stronger causal prior (Fig. 2). From equation (5), the perceptual binding in the action-outcome interval is given by $(\tau_O - \tau_A - \mu_{AO})(\sigma_A^2 + \sigma_O^2)/\sigma_{tot}^2$ in the causal case ($\hat{\xi} = 1$) and none otherwise ($\hat{\xi} = 0$). Because the sensory signals are distributed according to $\tau_A \sim \mathcal{N}(t_A^* + d_A, \sigma_A^2)$ and $\tau_O \sim \mathcal{N}(t_O^* + d_O, \sigma_O^2)$, the average of $\tau_O - \tau_A - \mu_{AO}$ factor is $m = t_O^* - t_A^* + d_O - d_A - \mu_{AO}$. Hence, the sign of m determines if binding or repulsion is predicted on average. With the current set of parameters, m is positive in the voluntary condition, yielding binding, and negative in the involuntary condition, yielding repulsion (schematically drawn in Fig. 3a). Perceptual shift is almost zero regardless of the causal prior $P(\xi = 1)$ in the sham condition because $m \approx 0$. We chose $P(\xi = 1) = 0.1$ for this under-constrained sham condition, assuming that causality would not be frequently detected.

Our Bayesian model provide interesting insights on what possibly drives the perceived action-outcome temporal binding and repulsion effects. We empirically observed sensory delay d to increase with larger standard deviation σ of the Gaussian-distributed jitter (observed in both Haggard and Wolpe; see Table 1). This may imply that, as action or outcome ambiguity is increased due to noise (greater σ) for increased sensory uncertainty, more time would be needed (greater d) for a sensory input to reach the subject's perceptual threshold for temporal awareness in the baseline condition. Thus, because of the m's dependency on $d_O - d_A$, binding more likely happens when the outcome is unreliable (i.e., with large d_O) and repulsion more likely happens when the action is unreliable (i.e., with large d_A).

To further illustrate the model prediction from our simulations, we plotted separately the action and outcome perceptual shifts for the voluntary, involuntary and sham conditions as functions of the temporal disparity $\tau_O - \tau_A$ (c.f. equation (5)). Indeed, our data shows that for instances in which $\tau_O - \tau_A > \mu_{AO}$, action awareness is delayed (positive action shift, Fig. 3b) and outcome tone is anticipated (negative outcome shift, Fig. 3c), thereby demonstrating binding. The opposite happens when $\tau_O - \tau_A < \mu_{AO}$ thereby demonstrating repulsion in both action and outcome awareness (Figs. 3b and 3c, respectively). We then plotted how the model's MAP estimates on the action-outcome interval are affected by the sensory time difference $\tau_O - \tau_A$ in the baseline (here $\hat{\xi} = 0$ is forced; Fig. 3d) and operant (Fig. 3e) conditions. We observe from the baseline condition that the MAP estimates follow sensory inputs, $\hat{t}_O - \hat{t}_A = \tau_O - \tau_A$, whereas the perception of action and outcome timings shifted towards the prior mean, $\hat{t}_O - \hat{t}_A \approx \mu_{AO}$, in the voluntary and involuntary conditions but not so much in the sham condition with weak causal prior. Therefore, our model is agnostic as to whether the action is self-intended or unintended. Binding towards $\hat{t}_O - \hat{t}_A \approx \mu_{AO}$ will happen, be it in the opposite direction, as long as the action is believed to have caused the outcome. This suggests that causality is the phenomenon that underlies temporal binding, and likely SoA, with self-intended causality being a specific case. The temporal window of $\tau_O - \tau_A$ for detecting causality is wider in the voluntary and involuntary conditions than the sham condition.

We then examined how the prior belief in causation affects our proposed measure for SoA in Haggard's experimental setup. Our model predicts CCE to strengthen together with the causal prior but its strength differs depending on the conditions even at the same strength of the prior (Fig. 4a). Interestingly, d_A and σ_A are the only parameters of our Bayesian model that differentiate the three conditions in this figure. As we described above, these two parameters are empirically correlated such that the delay d_A increases with larger σ_A. Hence, the difference in CCE in the three conditions can be attributed to the inequalities in the standard deviations of the subjects' action timing estimation errors in the three conditions: $\sigma_A^{Vol} < \sigma_A^{Sham} < \sigma_A^{Invol}$ as per Haggard et al.'s data. Haggard et al. speculated that the unexpected and surprising quality of the TMS-induced movement could account for the repulsion effect in the involuntary condition. We suggest that this surprise might have introduced uncertainty in the perception of action input signals. Hence, while subjects were certain of the nature of their voluntary actions, they could be less certain of the proprioception signals induced by TMS, which could explain the inequalities in σ_A. As a result, the model gives $CCE^{Vol} > CCE^{Sham} > CCE^{Invol}$ according to the requirement (C), i.e., reliable sensory inputs, for having high CCE when compared at the same strength of the causal prior.

The relation between CCE and SoA becomes clear when we analyze them with the fitted values of the causal prior ($P(\xi = 1) = 0.9$ for the voluntary and involuntary conditions and $P(\xi = 1) = 0.1$ for the sham condition as indicated in Table 1). Figure 4b plots CCE on a per trial basis as functions of the temporal disparity $\tau_O - \tau_A$ (c.f. the analytical expression for CCE in Methods). CCE in the voluntary condition has a higher peak than the involuntary condition as we described above (due to small σ_A in the voluntary condition for the requirement (C)). In

both voluntary and involuntary conditions, CCE diminishes as $\tau_O - \tau_A$ moves farther from μ_{AO} because of the requirement (A) of small $|\tau_O - \tau_A - \mu_{AO}|$ for having high CCE. Finally, CCE for the sham condition takes much lower values than the voluntary or involuntary conditions because of the requirement (B) of large $P(\xi = 1)$ for having high CCE. Hence, our Bayesian model coherently explains not just SoA that arises from the causation of the outcome by the action, but also the one that is influenced by the reliability of the different agency cues – a *precision*-dependent *causal agency*.

In a similar fashion, we then examined the underlying psychophysical mechanisms that could account for the temporal binding observed by Wolpe and colleagues, in which three uncertainty levels (high, intermediate, and low uncertainty) of the outcome stimulus were tested. We use the Bayesian model that was used to reproduce the Haggard's experiments with the same values of μ_{AO} and σ_{AO} but adjusted the strength of the causal prior $P(\xi = 1)$ to fit the reported action timing and outcome timing in each condition. We used $P(\xi = 1) = 0.9$, 0.6 and 0.5 for low, intermediate and high tone uncertainty conditions, respectively (see Table 1 and Methods). This means that the prior belief in causation decreases with the tone uncertainty, which is plausible.

Our model reproduces the experiments of Wolpe et al. (Fig. 5a), qualitatively explaining the temporal binding they observed in terms of a single, coherent cue integration formulation. The Bayesian estimate of the action-outcome intervals shift towards $\hat{t}_O - \hat{t}_A \approx \mu_{AO}$, as per the causal temporal prior in equation (2) when causality is detected. On the one hand, the magnitude of the shift is greater when the outcome uncertainty is high (c.f. equation (5)). But, on the other hand, causality is less frequently detected when the outcome uncertainty is high with the reduced causal prior. These two opposing effects are summarized in Fig. 5b. The model can qualitatively reproduce the experiments if the former effect is more dominant. Quantitatively, however, the latter effect is necessary to mitigate the former effect.

Next, we plot how the Bayesian estimate of the action-outcome interval, $\hat{t}_O - \hat{t}_A$, depends on the sensory inputs, $\tau_O - \tau_A$. The perceived intervals faithfully follow the sensory inputs in the baseline condition (Fig. 5c), where all trials are acausal ($\hat{\xi} = 0$) by definition. In the operant condition (Fig. 5d), the Bayesian estimate shifts toward the prior assumption $\hat{t}_O - \hat{t}_A \approx \mu_{AO}$ when the sensory inputs are highly consistent with the prior $\tau_O - \tau_A \approx \mu_{AO}$ and, thus, when the causality is detected ($\hat{\xi} = 1$). Otherwise, the estimate of action-outcome intervals follows sensory inputs. The temporal window of $\tau_O - \tau_A$ for detecting causality is wider when the outcome uncertainty is lower.

Next, we quantify again CCE as a possible measure of SoA. CCE diminishes with outcome uncertainty even when compared at the same level of causal prior (Fig. 6a). Hence, CCE explicitly depends on the outcome uncertainty. When plotted as functions of temporal disparity, with the specific causal priors obtained for each outcome uncertainty condition, the peak values of CCE noticeably differ across the uncertainty conditions (Fig. 6b). This is because of the different values of the outcome uncertainty σ_O but also partly because of the different values of the causal prior. In all conditions, CCE falls off with the disparity of sensory inputs from the prior mean, $|\tau_O - \tau_A - \mu_{AO}|$. This fall off is milder when the uncertainty is lower. These results clearly manifest again three basic requirements of CCE: (A) the consistency of sensory inputs with the causal prior; (B) strong prior belief in causality; and (C) reliable sensory inputs.

Discussion

We formalize SoA by drawing parallels from a Bayesian inference of the *ventriloquism effect* that estimates a common cause behind its multisensory integration. Our Bayesian model integrates the action-outcome signals, compares them with the prior expectation, and infers the causality between them as well as the timing of these sensory signals. Our model could concisely reproduce the intentional binding experiments by Haggard et al.'s and Wolpe et al. The temporal binding or repulsion phenomena are explained by the compromise between the noisy sensory observations and the prior belief of the action-outcome timing. Importantly, our Bayesian model predicts that the perceptual binding is generally trial-dependent, and it must be correlated with the estimated causality $\hat{\xi}$ between the action and outcome. This prediction can be tested, when the probability P_c for detecting causality is not close to 0 or 1, by examining if the distribution of action-outcome intervals is bimodal and if the intervals correlate with the reported causality between the action and outcome. In this work, we focus on the timing to investigate the intentional binding effects but the mathematical elucidations of our model can permit other modalities (e.g., visual or haptic) and structural properties (e.g., inter alia, location, size, shape and texture).

In addition, we theorize SoA as *the confidence in causal estimate*: CCE. CCE is high when the action-outcome timing is consistent with the causal prior, the causal prior is strong, and the action and outcome signals are reliable. This notion is consistent to what have been propounded as demonstrations of SoA: SoA arises from the causal relation between performed actions and their consequences[4,19,22,23] and from the integration of different agency cues whose individual influences are determined by their reliability[14,15,24-27]. Hence, we posit CCE to be a plausible measure of SoA.

Specifically, we postulate CCE fits the notion of a pre-reflective, implicit *feeling of agency* (FoA). Synofzik and colleagues[4,25] provide a compelling account of such feeling: FoA is best accounted for by multimodal weighting and integration of different agency cues, and consists of an automatic registration of whether an action or sensory event is caused by the self or not. They posit FoA is nothing other than first-person in that the self is implied, hence, no external attribution (e.g., to TMS that caused the action) is possible. In the event that there is a feeling of exogenous causation, this will be overwritten by an explicit, interpretative judgment of agency (JoA) based on contextual beliefs or rationalizations. Similarly, the analytical expression of CCE shows that it is a multimodal weighting and integration process that lies at the center of obtaining a Bayesian causality inference. Furthermore, CCE itself does not attribute causality to any external agent, such as in the case of strong causal prior for TMS-induced movements. The judgement of the causality, $\hat{\xi}$, is then made based on the posterior ratio r that compares CCE with the confidence in the acausal estimate. Perceptual timing in our model simply reflects the sensory signals if the causality is not detected ($\hat{\xi} = 0$), whereas they are overwritten by the influence of the prior if the causality is detected ($\hat{\xi} = 1$). For example, in the involuntary condition of Haggard et al., the estimated action and outcome timing by the model repulse reflecting the judgment of the causality. A compelling speculation in Haggard et al.'s paper[6] suggests this notion: the repulsion in the involuntary condition "reflects a mental operation to segregate, and thus to discriminate, pairs of events that cannot plausibly be linked by our own causal agency" (p. 384). We suggest such mental operation fits the notion of JoA, as quantified by the time shifts in equation (5) with the detected causality, and the peculiar feeling of causation by the involuntary movement to be FoA, quantified by CCE.

Following the above explanation, our theory therefore has a different take of Haggard et al.'s binding effect that requires intentionality. CCE argues that the judgment of the causality is central to the perceived temporal binding, consistent with current evidence that competes with the intentional account: the temporal binding is actually *causal*, not intentional[22]. For

example, our model judges the causation of the tone even by the TMS-induced action in the involuntary condition. Hence, our Bayesian model predicts this unintended causality. Furthermore, our Bayesian model predicts that the action-outcome timing shifts toward the prior belief, $\hat{t}_O - \hat{t}_A \approx \mu_{AO}$, when the causality is perceived irrespective of the nature of the action, whether self-generated (i.e., the voluntary condition) or unintended (i.e., the involuntary condition). We suggest that the repulsion happened in the involuntary condition because of the proprioceptive noise (large σ_A) that characterizes the TMS. The large proprioceptive noise may be caused by the internal prediction error resulting from unintended proprioceptive signal[28,12] or the large clicking sound of TMS. In this sense, intentionality is just one factor that influences CCE and perceptional shift in our model. We predict that an experimental manipulations that reduces σ_A would increase perceived SoA even for unintended artificial actions. The prediction is therefore distinct from what was previously considered and can therefore serve as testable prediction for future experiments on causal agency.

Our theory also has a different take of Wolpe et al.'s binding effect. Wolpe and colleagues showed intentional binding as cue integration with uncertainty in outcome signals. They speculated that action and outcome bindings are driven by two distinct mechanisms: action binding is predicted by cue integration, but outcome binding supports the predictive pre-activation hypothesis[29], i.e., the neural representation of the sensory outcome is activated prior to it. Hence, the outcome signals are perceived faster with less jitter than when it is not predicted to occur after the action. This could explain why the subjects' timing estimations are largely erroneous in the baseline condition and why the outcome binding is greater than the action binding. Our theory, although qualitative, explains both action and outcome bindings by a single Bayesian cue integration mechanism. Our model explains that the magnitudes of the action and outcome perceptual shifts, $(\tau_O - \tau_A - \mu_{AO})$, are influenced primarily by the ambiguity of the outcome sensory signals, $(\sigma_A^2 + \sigma_O^2)/\sigma_{tot}^2$, and also in part by the strength of the causal prior that diminishes with outcome uncertainty.

The intentional binding paradigm has also been used to study pathological sense of agency[30-32]. Patients with schizophrenia tend to have much stronger temporal binding than healthy volunteers. Moreover, unlike healthy volunteers, their temporal binding of action timing does not depend on the probability of the outcome tone presentation[32]. These results are explained by our Bayesian model by assuming that schizophrenia patients cannot easily adapt their abnormally strong belief in causality (i.e., too large $P(\xi = 1)$) and the uncertainty in the outcome (i.e., σ_O). Another important point is that, unlike healthy volunteers, patients with schizophrenia exhibit temporal binding of action timing that depends on the presence or absence of the outcome. It will be an interesting future study to model this result by explicitly incorporating the probabilistic occurrence of the outcome in our Bayesian model.

In summary, we posit that since the Bayesian cue integration is primarily precision-dependent so is our theory of SoA. Our model predicts that if the uncertainty of the sensory input signals could be maintained small, even unintended causal action may give rise to high CCE (hence, strong SoA) – hence, our notion of precision-dependent casual agency. We posited the precise estimation that gives rise to SoA encapsulates consistency in the perceived action-outcome effect, the prior belief of the causation of the outcome by the action, and the reliability of the perceived sensory signals. This theory may shed light on the mechanism of reduced SoA in psychosis, the understanding of the difference between FoA and JoA, and the design of prosthetic devises that heighten SoA.

Acknowledgement
This study was supported by Brain/MINDS from AMED under Grant Number JP18dm020700 and JSPS KAKENHI Grant Number JP18H05432.

Methods

Analytical Expressions for the Bayesian Estimates

The MAP estimate (equation (3)) of the Bayesian observer has a simple analytical expression. The MAP estimation is computed based on the posterior probability $P(t_A, t_O, \xi|\tau_A, \tau_O) = P(\tau_A, \tau_O, t_A, t_O, \xi)/P(\tau_A, \tau_O)$, where the peak location only depends on the joint distribution $P(\tau_A, \tau_O, t_A, t_O, \xi)$ in the numerator. The joint distribution is decomposed as $P(\tau_A, \tau_O, t_A, t_O, \xi) = P(\tau_A|t_A)P(\tau_O|t_O)P(t_A, t_O|\xi)P(\xi)$, where the conditional distributions for action and outcome are $P(\tau_A|t_A) = \exp[-(t_A - \tau_A)^2/(2\sigma_A^2)]/\sqrt{2\pi\sigma_A^2}$ and $P(\tau_O|t_O) = \exp[-(t_O - \tau_O)^2/(2\sigma_O^2)]/\sqrt{2\pi\sigma_O^2}$, respectively, and the prior distribution is

$$P(t_A, t_O|\xi) = \begin{cases} \exp\left(-\frac{(t_O - t_A - \mu_{AO})^2}{2\sigma_{AO}^2}\right)/Z_1 & (\xi = 1) \\ 1/Z_0 & (\xi = 0) \end{cases}$$

with normalization constants $Z_1 \equiv \int_R \exp\left(-\frac{(t_O - t_A - \mu_{AO})^2}{2\sigma_{AO}^2}\right) dt_A dt_O \approx \sqrt{2\pi}\sigma_{AO}T$ and $Z_0 \equiv \int_R dt_A dt_O = T^2$.

We separately compute the peak location $(\hat{t}_A, \hat{t}_O)$ for the causal case $\xi = 0$ and the acausal case $\xi = 0$ and, then, compare these two peaks. In the acausal case, because $P(\tau_A|t_A)$ and $P(\tau_O|t_O)$ take the maximum values at $t_A = \tau_A$ and $t_O = \tau_O$, respectively, the location of the acausal peak is $(\hat{t}_A, \hat{t}_O)|_{\xi=0} = (\tau_A, \tau_O)$ and the peak value is $\max_{t_A, t_O} P(\tau_A, \tau_O, t_A, t_O, \xi = 0) = \frac{P(\xi=0)}{2\pi\sigma_A\sigma_O Z_0}$. In the causal case, the peak of the joint distribution is found by minimizing a quadratic function. The peak location is given by $(\hat{t}_A, \hat{t}_O)|_{\xi=1} = \left(\tau_A + \frac{\sigma_A^2}{\sigma_{tot}^2}(\tau_O - \tau_A - \mu_{AO}), \tau_O - \frac{\sigma_O^2}{\sigma_{tot}^2}(\tau_O - \tau_A - \mu_{AO})\right)$, where $\sigma_{tot}^2 \equiv \sigma_A^2 + \sigma_O^2 + \sigma_{AO}^2$ is the total variance, and the peak value is computed as $\max_{t_A, t_O} P(\tau_A, \tau_O, t_A, t_O, \xi = 1) = \frac{P(\xi=1)}{2\pi\sigma_A\sigma_O Z_1}\exp\left(-\frac{(\tau_O - \tau_A - \mu_{AO})^2}{2\sigma_{tot}^2}\right)$. We define the log-ratio of the posterior peaks for $\xi = 1$ and $\xi = 0$ by

$$r \equiv \frac{\max_{t_A, t_O} P(t_A, t_O, \xi = 1|\tau_A, \tau_O)}{\max_{t_A, t_O} P(t_A, t_O, \xi = 0|\tau_A, \tau_O)} = \exp\left(\theta - \frac{(\tau_O - \tau_A - \mu_{AO})^2}{2\sigma_{tot}^2}\right)$$

with $\theta \equiv \log\left[\frac{P(\xi=1)Z_0}{P(\xi=0)Z_1}\right]$. If $r > 1$, the MAP estimate is given by $(\hat{t}_A, \hat{t}_O)|_{\hat{\xi}=1}$ and $\hat{\xi} = 1$, which predicts perceptual shifts. If $r < 1$, the MAP estimate is given by $(\hat{t}_A, \hat{t}_O)|_{\hat{\xi}=0}$ and $\hat{\xi} = 0$, which predicts no perceptual shifts. The probability for detecting causality (i.e., $\hat{\xi} = 1$) is also easily computable because $\tau_O - \tau_A - \mu_{AO}$ is distributed according to the Gaussian distribution $\mathcal{N}(m, \sigma_A^2 + \sigma_O^2)$ with $m \equiv t_O^* - t_A^* + d_O - d_A - \mu_{AO}$. Hence, the causality is detected if $|\tau_O - \tau_A - \mu_{AO}| < \sqrt{2\theta}\sigma_{tot}$, and this happens with probability

$$P_c = \frac{1}{2}\left[\mathrm{erf}\left(\frac{\sqrt{2\theta}\sigma_{tot} - m}{\sqrt{2(\sigma_A^2 + \sigma_O^2)}}\right) + \mathrm{erf}\left(\frac{\sqrt{2\theta}\sigma_{tot} + m}{\sqrt{2(\sigma_A^2 + \sigma_O^2)}}\right)\right]$$

Next, we evaluate the confidence in the causal MAP estimation $CCE \equiv \max_{t_A, t_O} P(t_A, t_O, \xi = 1|\tau_A, \tau_O)$, which comprises the numerator of the ratio r. To quantify this confidence, we need to first evaluate $P(\tau_A, \tau_O) = P(\tau_A, \tau_O, \xi = 1) + P(\tau_A, \tau_O, \xi = 0)$ with

$$P(\tau_A, \tau_O, \xi = 1) = \iint_R P(t_A, t_O, \xi = 1, \tau_A, \tau_O) dt_A dt_O = \frac{P(\xi=1)\sigma_{AO}}{Z_1\sigma_{tot}}\exp\left(-\frac{(\tau_O - \tau_A - \mu_{AO})^2}{2\sigma_{tot}^2}\right)$$

and

$$P(\tau_A, \tau_O, \xi = 0) = \iint_R P(t_A, t_O, \xi = 0, \tau_A, \tau_O) dt_A dt_O = \frac{P(\xi=0)}{Z_0}$$

Combining these expressions together, we obtain

$$CCE = \max_{t_A, t_O} P(\tau_A, \tau_O, t_A, t_O, \xi = 1)/P(\tau_A, \tau_O)$$

$$= \frac{\sigma_{tot}}{2\pi\sigma_A\sigma_O\sigma_{AO}} \text{Sigmoid}\left(\theta - \frac{(\tau_O - \tau_A - \mu_{AO})^2}{2\sigma_{tot}^2} + \log\frac{\sigma_{AO}}{\sigma_{tot}}\right)$$

where $\text{Sigmoid}(x) = 1/(1 + e^{-x})$ is the sigmoid function.

Validation of Bayesian Estimates

The principal measure of intentional binding is the mean perceptual shift of temporal awareness of action and sensory outcome. A *perceptual shift* is the change in the subjective estimation of action or outcome timing from the baseline (*Bsln*) to the operant (*Oprnt*) condition, which is computed as

$$\delta_A = E\left[\hat{t}_A^{Oprnt}\right] - E\left[\hat{t}_A^{Bsln}\right]$$
$$\delta_O = E\left[\hat{t}_O^{Oprnt}\right] - E\left[\hat{t}_O^{Bsln}\right].$$

A *positive* shift informs the perception of timing shifted later in time, and a *negative* shift informs the perception of timing shifted earlier in time.

With the mean action and outcome perceptual shifts, δ_A and δ_O, we can obtain the *model estimation error*, i.e., the difference of our model's (*Model*) prediction of the action or outcome perceptual shift to the corresponding perceptual shift reported in the experiments (*Exp*). We obtain this as

$$Err_A = \left|\delta_A^{Model} - \delta_A^{Exp}\right|$$
$$Err_O = \left|\delta_O^{Model} - \delta_O^{Exp}\right|.$$

Simulation Details

Table 1 lists all the parameters of our Bayesian model. We performed different simulations to reproduce the action and outcome perceptual shifts that were reported in the experiments, and explain their underlying psychophysical mechanisms in Bayesian terms. We generated 35,000 instances of τ_A and τ_O pairs for each experimental condition. To determine the optimal model parameter values, we performed *model fitting*: (*i*) a set of possible model parameter values are used to simulate and obtain a prediction dataset, (*ii*) the model estimation error (explained above) is computed to determine the difference of our model's predictions to the reported empirical results, and (*iii*) the model parameters that best minimized this error are selected.

Simulations 1| The objective was to determine the values of the operant parameters μ_{AO} and σ_{AO} that best estimate the results reported by Haggard, Clark & Kalogeras[6]. We tested using the baseline parameters in Table 1-Set A and different combinations of μ_{AO} and σ_{AO}. We assumed initially that the action always causes the outcome, i.e., $P(\xi = 1) = 1$. We later dropped this assumption to test the effect of different causal priors.

For each μ_{AO} and σ_{AO} pair, we obtained the model estimation errors for the reported action and outcome perceptual shifts listed in Table 2-Set A. We took the average of the model estimation errors for the voluntary, involuntary and sham conditions to obtain a single model estimation error. We looked at the model estimation errors for the (a) action perceptual shifts only, (b) outcome perceptual shifts only, and (c) action-outcome perceptual shifts. After testing different μ_{AO} and σ_{AO} pairs, our results showed the best estimates of the model to be at $\mu_{AO} = 230$ and $\sigma_{AO} = 10$. Furthermore, we observed the perceptual shift in action timing alone was sufficient to indicate the best estimates of the model.

Simulations 2| The objective was to obtain the causal prior probability that yield the best estimates of Haggard et al.'s results. With $\mu_{AO} = 230$ and $\sigma_{AO} = 10$, we tested for $P(\xi = 1)$ in the range 0 to 1 with increments of 0.1. We used the same pairs of τ_A and τ_O from simulations (1), together with the Table 1-Set A baseline parameters, when we carried out our simulations. We computed once again the model estimation errors for the empirical results listed in Table 2–Set A. We selected the $P(\xi = 1)$ that best minimized the estimation errors for the voluntary, involuntary and sham conditions. Table 1-Set C includes the parameters that yielded the best model estimates. Fig. 1 shows the action and outcome perceptual shifts, as well as the intervals between perceptual shifts, which were obtained by our Bayesian model using these parameters.

Simulations 3| The objective was to reproduce the perceptual shifts reported by Wolpe, Haggard, Siebner & Rowe[20], listed in Table 2-Set B. We generated another set of 35,000 τ_A and τ_O pairs using, this time, the baseline parameters listed in Table 1–set B. We performed simulations and tested using $\mu_{AO} = 230$, $\sigma_{AO} = 10$, and $P(\xi = 1)$ in the range 0 to 1 with increments of 0.1. We did not perform additional simulations to redetermine μ_{AO} and σ_{AO} since our aim is to reproduce qualitatively all the experiments with the same μ_{AO} and σ_{AO} as possible in order to have simple yet consistent explanations by our Bayesian model. Although we did not modify here μ_{AO} and σ_{AO}, our analyses and results can show that their effects can be predicted and explained by our model. The model estimation errors once again indicate the estimates of action perceptual shifts led to the best estimates of the model. We list under Table 1-Set D the $P(\xi = 1)$ that yielded the best estimates of the model for the low, intermediate and high uncertainty tone conditions. We show in Fig. 5a the action and outcome perceptual shifts, and intervals between shifts, predicted by our Bayesian model for this experimental setup.

Simulations 4| The objective was to determine the influence of the causal prior and the temporal difference $\tau_O - \tau_A$ (that varies in every trial) on the various predictions of our Bayesian model for Haggard et al.'s experimental setup. We used the model parameters and τ_A and τ_O pairs from simulations (1) and (2). We obtained our Bayesian model's predictions of the intervals between action and outcome perceptual shifts, binding and repulsion effects, action-outcome timing interval, $\hat{t}_O - \hat{t}_A$, in the baseline and operant conditions, and CCE. The results are shown in Figs. 2, 3 and 4.

Simulations 5| The objective and target results were the same as simulations (4), but using this time the model parameters and τ_A and τ_O pairs from simulations (3) to account for Wolpe et al.'s experimental setup. The resulting plots are shown in Figs. 5 and 6.

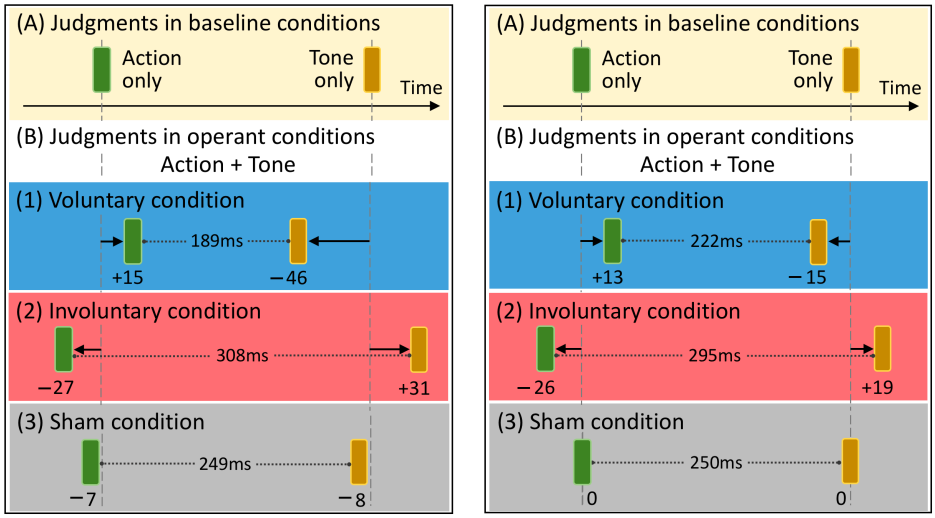
Table 1. List of Bayesian model parameters and their values

Baseline condition parameters	d_A	σ_A	d_O	σ_O
	(unit is ms)			
• *Set A*: Reported by Haggard, Clark & Kalogeras[6]				
Voluntary action	6	66		
Involuntary action (TMS-induced muscle twitch)	83	83		
Sham TMS (audible click only)	32	78		
Auditory tone			15	72
• *Set B:* Reported by Wolpe, Haggard, Siebner & Rowe[20]				
Voluntary action	-8	75		
Low tone uncertainty			35	61
Intermediate tone uncertainty			46	66
High tone uncertainty			95	90

Operant condition parameters	$\mu_{AO} = 230$ (unit is ms)	$\sigma_{AO} = 10$	$P(\xi = 1)$
• *Set C*: Obtained by our Bayesian model			
Voluntary condition			0.9
Involuntary condition			0.9
Sham condition			0.1
• *Set D*: Obtained by our Bayesian model			
Low uncertainty tone condition			0.9
Intermediate uncertainty tone condition			0.6
High uncertainty tone condition			0.5

Table 2. Reported perceptual shifts in action and outcome temporal awareness

Judged event (operant condition)	*Mean estimation error* (ms)	*Mean perceptual shift* (ms)
• *Set A*: Reported by Haggard, Clark & Kalogeras[6]		
Voluntary action, then	21	15
tone	-31	-46
Involuntary action, then	56	-27
tone	46	31
Sham TMS, then	25	-7
tone	7	-8
• *Set B*: Reported by Wolpe, Haggard, Siebner & Rowe[20]		
Action, then	31	39
low uncertainty tone	-16	-51
Action, then	23	31
intermediate uncertainty tone	-19	-65
Action, then	24	32
high uncertainty tone	-10	-105



Results from Haggard et al. (2002) Simulation results with Bayesian model

Figure 1| The qualitative replication of the empirical results reported by Haggard, Clark & Kalogeras[6] (left panel) by our Bayesian model (right panel). The subjects' mean judgment error for the single-event baseline condition is subtracted from the mean judgment error for the corresponding operant event, which results to the magnitude and direction of the perceptual shifts. A *positive* perceptual shift informs delayed awareness, and a *negative* shift informs anticipated awareness. The action and outcome timings are perceived to shift towards each other in the voluntary condition. In contrast, they are perceived to repulse in the involuntary condition. There is no discernible perceptual shift in the sham condition.

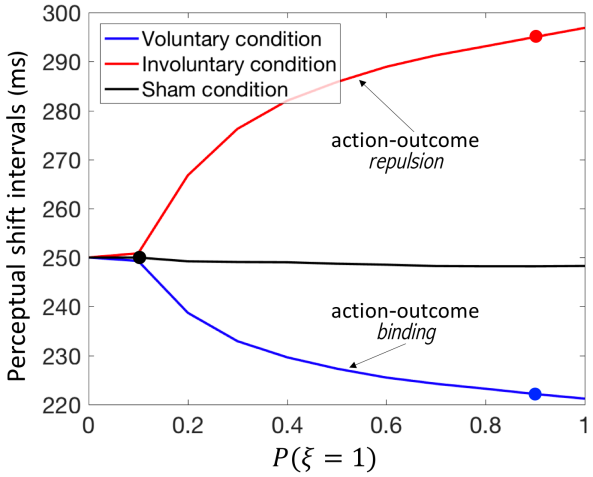


Figure 2| The Bayesian model predictions of the influence of causal prior strength on action-outcome perceptual shifts. The best estimates of the Bayesian model (in Fig. 1) were obtained from different causal priors, specifically, $P(\xi = 1)$ is 0.9, 0.9 and 0.1 (marked by the colored dots) for the voluntary, involuntary, and sham conditions, respectively. The intervals between the action and outcome perceptual shifts shrink in the voluntary, but widen in the involuntary, condition with a strong causal prior. Minimal changes in perceptual shifts are predicted for the sham condition even with a strong causal prior.

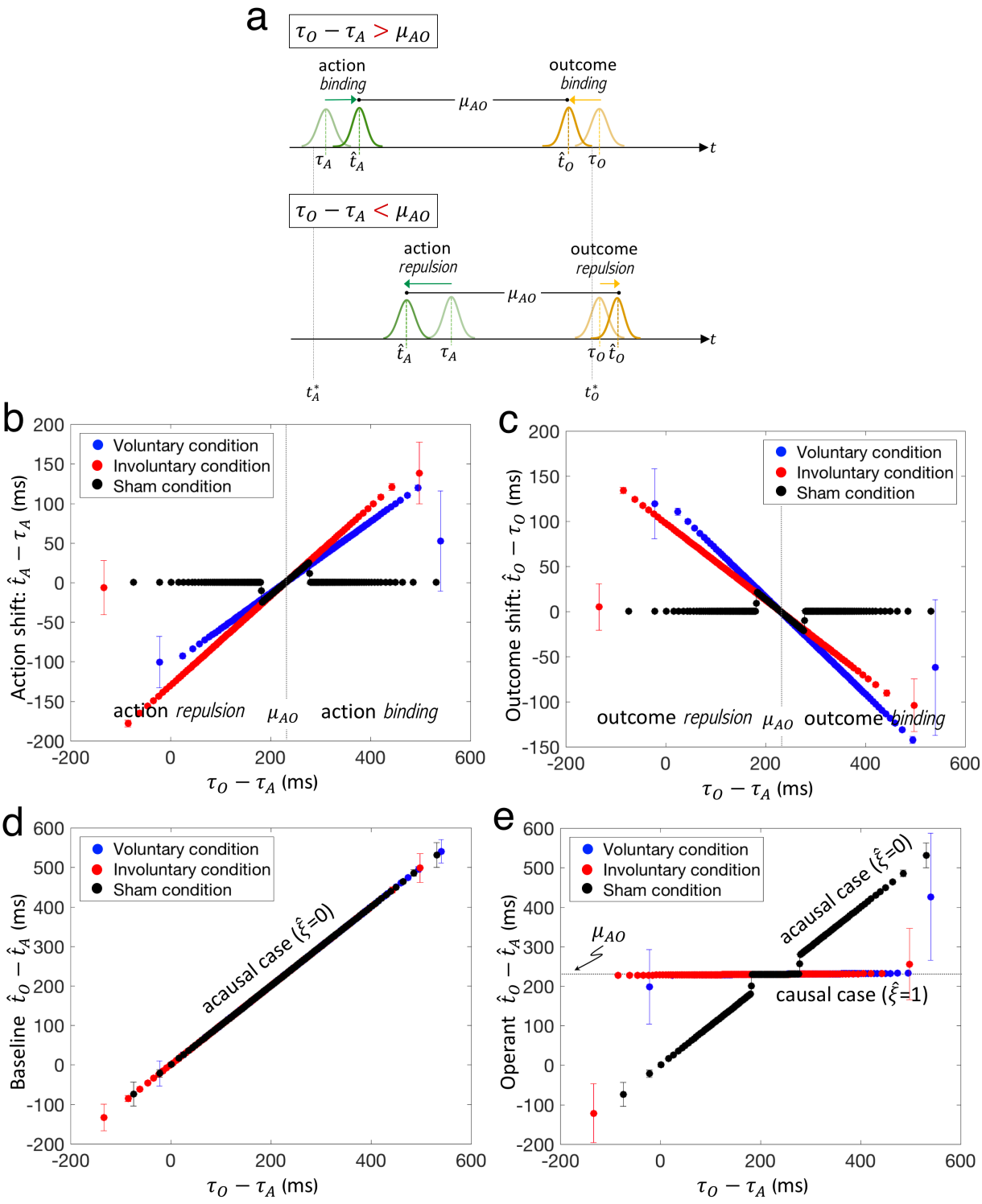


Figure 3| The Bayesian model predictions of trial-to-trial action-outcome temporal binding and repulsion effects, in **(b)** and **(c)**, and action-outcome timing interval, $\hat{t}_O - \hat{t}_A$, in the baseline and operant conditions, in **(d)** and **(e)**, as functions of the perceived temporal disparity, $\tau_O - \tau_A$. The causal prior $P(\xi = 1)$ is 0.9, 0.9 and 0.1 for the voluntary, involuntary, and sham conditions, respectively (as in Fig. 2). The per trial results are grouped accordingly into bins of width 200 (randomly chosen), and the mean and standard deviation for each bin are plotted. This format is followed each time a quantity of interest is plotted as a function of $\tau_O - \tau_A$. **a**| Our Bayesian model predicts that (shown schematically) if $\tau_O - \tau_A > \mu_{AO}$ action and outcome

binding will happen. Otherwise, i.e., $\tau_O - \tau_A < \mu_{AO}$, action-outcome repulsion will occur. In both cases, the perceived timings in the baseline move (compress or stretch) towards the temporal consistency $\hat{t}_O - \hat{t}_A \approx \mu_{AO}$ in the operant condition. **b, c**| When $\tau_O - \tau_A > \mu_{AO}$, there is positive perceptual shift in action awareness ($\hat{t}_A - \tau_A > 0$) and negative perceptual shift in outcome awareness ($\hat{t}_O - \tau_O < 0$). The opposite happens when $\tau_O - \tau_A < \mu_{AO}$. *Both* binding and repulsion occur in *both* voluntary and involuntary conditions, but very little effect in the sham condition. **d**| The Bayesian estimates follow the sensory inputs in the baseline condition, i.e., $\tau_O - \tau_A \approx \hat{t}_O - \hat{t}_A$, where all trials are acausal ($\hat{\xi} = 0$) by definition. **e**| The Bayesian estimate shifts toward the prior assumption, $\hat{t}_O - \hat{t}_A \approx \mu_{AO}$, when the sensory inputs are highly consistent with the prior, $\tau_O - \tau_A \approx \mu_{AO}$, and therefore when causality is detected ($\hat{\xi} = 1$). Otherwise, the estimate of action and outcome timings follow the sensory inputs.

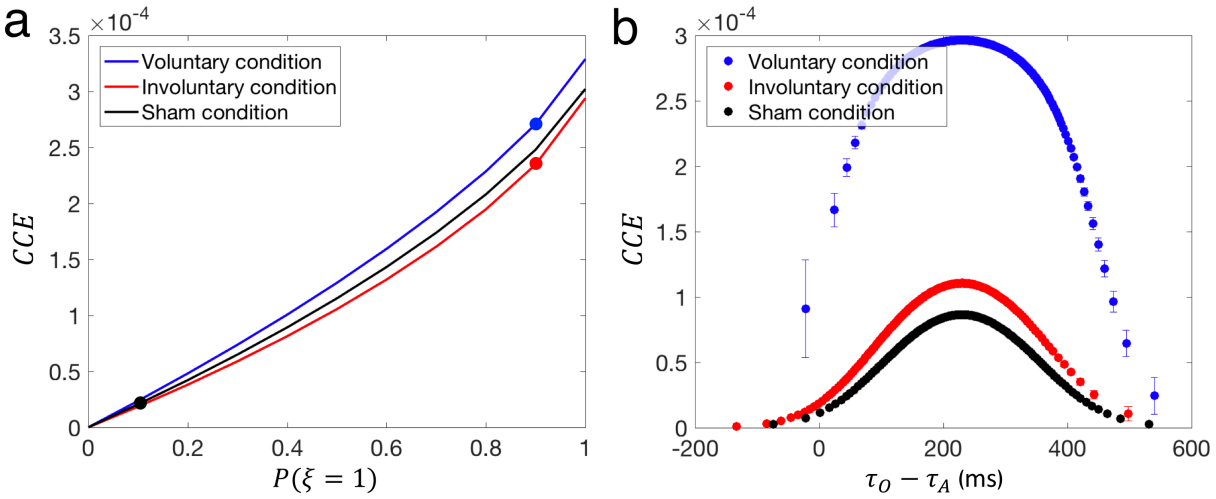


Figure 4| The Bayesian model predictions of CCE, i.e., the *confidence in causal estimate*, which is our proposed measure for SoA. **a**| Our Bayesian model predicts CCE to increase with a stronger causal prior. Furthermore, CCE differs for each condition even with equal prior strengths. This can be attributed to the difference in the amplitude of the jitter in the self-generated versus TMS-induced (muscle twitches and audible clicks) movement. **b**| When plotted as functions of the trial-to-trial temporal disparity $\tau_O - \tau_A$, with the specific causal priors obtained for each condition, marked in **(a)**, CCE has a higher peak in the voluntary condition, but much lower values in the sham condition. Furthermore, CCE diminishes as the temporal disparity in sensory inputs moves further away from the prior mean, $|\tau_O - \tau_A - \mu_{AO}|$. This falling of the CCE is faster when the causal prior is weaker and the uncertainty in the action input signal is higher.

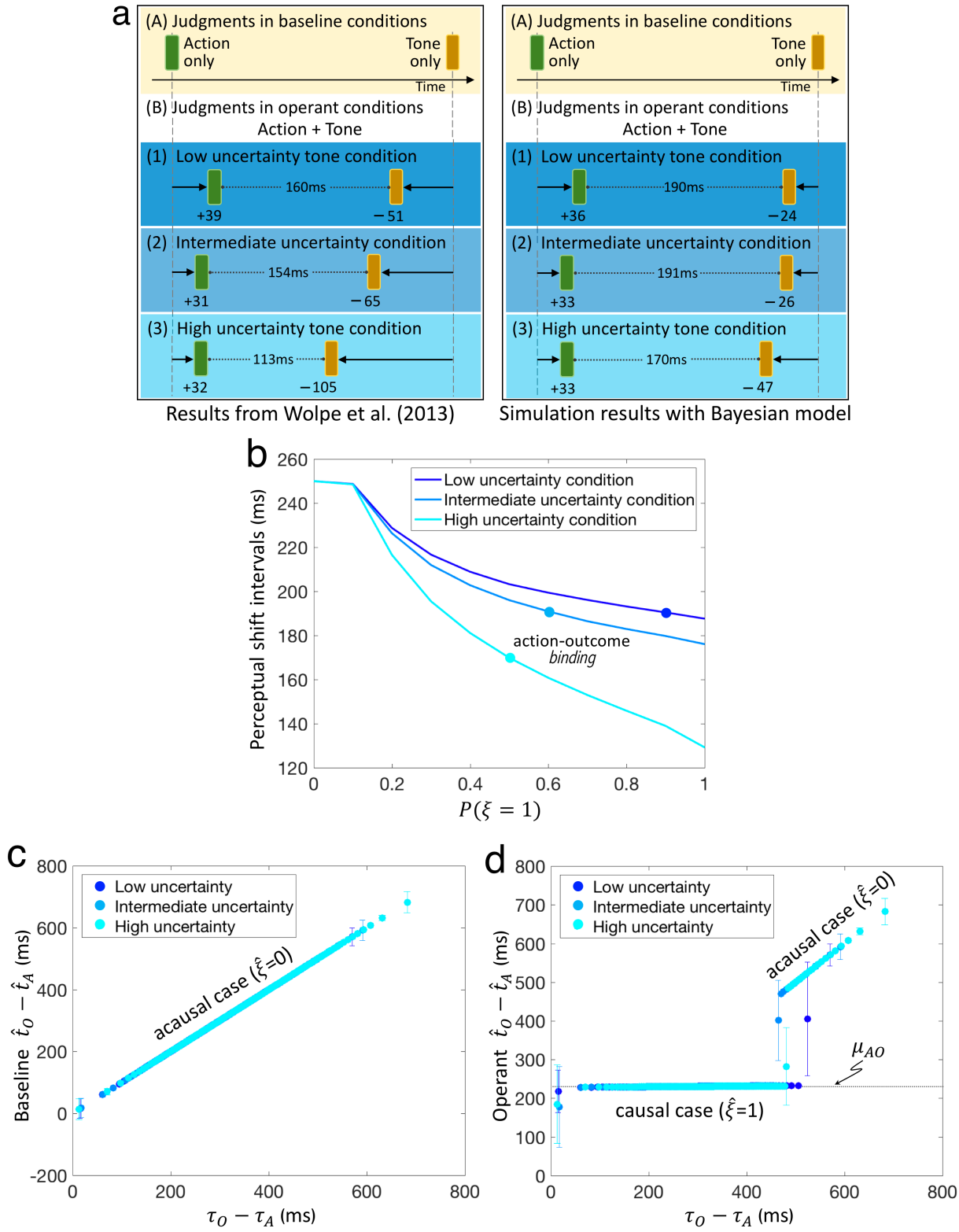


Figure 5| The Bayesian model qualitative replications of, as well as predictions related to, the results reported by Wolpe, Haggard, Siebner & Rowe[20]. **a**| Qualitative replication of the experimental results (left panel) by our Bayesian model (right panel). **b**| The action-outcome binding increases under heightened uncertainty. However, causality is less detected when the causal prior is lower, which decreases the action-outcome binding effect. The best estimates of the Bayesian model in **(a)** were obtained from different causal prior strengths, specifically, $P(\xi = 1)$ is 0.9, 0.6 and 0.5 (marked by the colored dots) for the low, intermediate and high

tone uncertainty conditions, respectively. **c, d**| The causal prior strengths that correspond to each condition were used for the Bayesian estimate of the action-outcome timing interval $\hat{t}_O - \hat{t}_A$ in the baseline and operant conditions. The Bayesian estimate follows the sensory inputs in the baseline condition where all trials are acausal, but shifts toward the prior assumption, $\tau_O - \tau_A \approx \mu_{AO}$, when causality is detected. The temporal window of $\tau_O - \tau_A$ for detecting causality is wider when the outcome uncertainty is lower, which means more instances demonstrate binding.

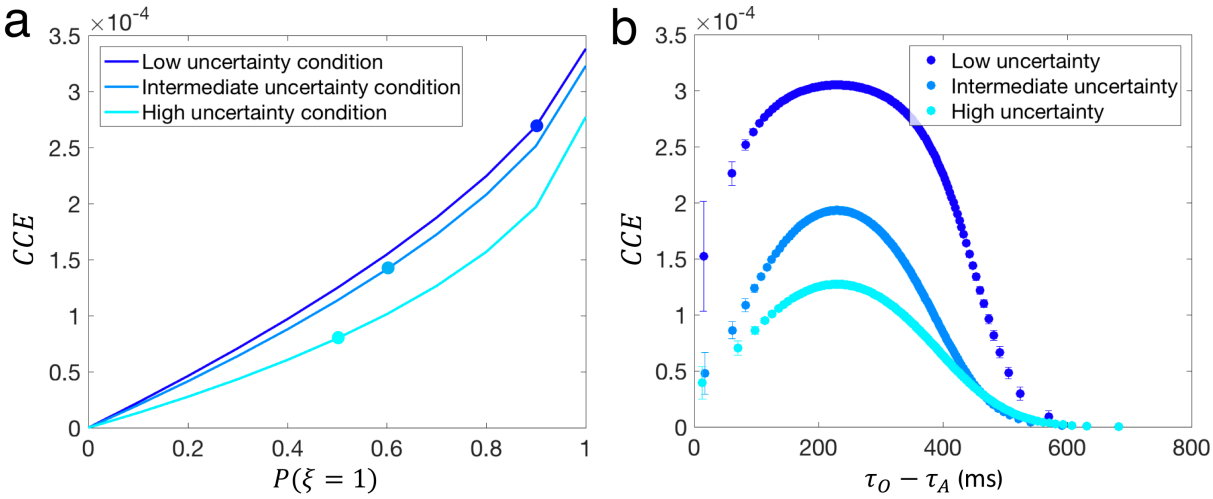


Figure 6| The Bayesian model prediction of CCE as a function of **(a)** the causal prior and **(b)** the temporal disparity $\tau_O - \tau_A$. **a**| The different effects of the causal prior on CCE across the three conditions is evident even with equal causal priors, which means that CCE depends on outcome uncertainty. **b**| When plotted as functions of the temporal disparity $\tau_O - \tau_A$, given the condition-dependent causal priors (marked by the colored dots in **(a)**), CCE falls off with the disparity of sensory inputs from the prior mean, $|\tau_O - \tau_A - \mu_{AO}|$, faster when the outcome uncertainty is higher.

1. David, N., Obhi, S. & Moore, J.W. Editorial: Sense of agency: examining awareness of the acting self. *Front. Hum. Neurosci.* **9**:310 (2015). https://doi.org/10.3389/fnhum.2015.00310
2. Moore, J.W. What is the sense of agency and why does it matter? *Front. Psychol.* **7**:1272 (2016). http://doi.org/10.3389/fpsyg.2016.01272
3. Gallagher, S. Multiple aspects in the sense of agency. *New Ideas in Psychology* **30(1)**, 15-31 (2012). https://doi.org/10.1016/j.newideapsych.2010.03.003
4. Synofzik, M., Vosgerau, G. & Newen, A. Beyond the comparator model: a multifactorial two-step account of agency. *Conscious. Cogn.* **17(1)**, 219-239 (2008). https://doi.org/10.1016/j.concog.2007.03.010
5. Limerick, H., Coyle, D. & Moore, J.W. The experience of agency in human-computer interactions: a review. *Front. Hum. Neurosci.* **8**:643 (2014). https://doi.org/10.3389/fnhum.2014.00643
6. Haggard, P., Clark, S. & Kalogeras, J. Voluntary action and conscious awareness. *Nat. Neurosci.* **5(4)**, 383-385 (2002). http://dx.doi.org/10.1038/nn827
7. Haggard, P. Human volition: towards a neuroscience of will. *Nat. Rev. Neurosci.* **9**, 934-946 (2008). http://dx.doi.org/10.1038/nrn2497
8. Frith, C. The psychology of volition. *Exp. Brain Res.* **229**, 289-299 (2013). https://doi.org/10.1007/s00221-013-3407-6
9. Haggard, P. Sense of agency in the human brain. *Nat. Rev. Neurosci.* **18**, 197-208 (2017). http://dx.doi.org/10.1038/nrn.2017.14
10. Gallagher, S. Philosophical conceptions of the self: implications for cognitive science. *Trends Cogn. Sci.* **4(1)**, 14-21 (2000). https://doi.org/10.1016/S1364-6613(99)01417-5
11. Wohlschläger, A., Haggard, P., Gesierich, B. & Prinz, W. The perceived onset time of self- and other-generated actions. *Psychol. Sci.* **14(6)**, 586-591 (2003). https://doi.org/10.1046/j.0956-7976.2003.psci_1469.x
12. David, N., Newen, A. & Vogeley, K. The "sense of agency" and its underlying cognitive and neural mechanisms. *Conscious. Cogn.* **17**, 523-534 (2008). https://doi.org/10.1016/j.concog.2008.03.004
13. Synofzik, M., Thier, P., Leube, D. T., Schlotterbeck, P. & Lindner, A. Misattributions of agency in schizophrenia are based on imprecise predictions about the sensory consequences of one's actions. *Brain* **133**, 262-271 (2010). https://doi.org/10.1093/brain/awp291
14. Moore, J.W. & Fletcher, P. C. Sense of agency in health and disease: a review of cue integration approaches. *Conscious. Cogn.* **21**, 59-68 (2012). https://doi.org/10.1016/j.concog.2011.08.010
15. Zalla, T. & Sperduti, M. The sense of agency in autism spectrum disorders: a disassociation between prospective and retrospective mechanisms. *Front. Psychol.* **6**:1278 (2015). https://doi.org/10.3389/fpsyg.2015.01278
16. Haggard, P. & Tsakiris, M. The experience of agency: feelings, judgments, and responsibility. *Current Directions in Psychological Science* **18**(4), 242-246 (2009). https://doi.org/10.1111/j.1467-8721.2009.01644.x
17. Caspar, E. A., Christensen, J. F., Cleeremans, A. & Haggard, P. Coercion changes the sense of agency in the human brain. *Curr. Biol.* **26**, 585-592 (2016). https://doi.org/10.1016/j.cub.2015.12.067
18. Sato, Y., Toyoizumi, T. & Aihara, K. Bayesian inference explains perception of unity and ventriloquism aftereffect: identification of common sources of audiovisual stimuli. *Neural Comput.* **19**, 3335-3355 (2007). https://doi.org/10.1162/neco.2007.19.12.3335
19. Moore, J. W. & Obhi, S. S. Intentional binding and the sense of agency: a review. *Conscious. Cogn.* **21(1)**, 546-561 (2012). https://doi.org/10.1016/j.concog.2011.12.002

20. Wolpe, N., Haggard, P., Siebner, H. R. & Rowe, J. B. Cue integration and the perception of action in intentional binding. *Exp. Brain Res.* **229**, 467-474 (2013). https://doi.org/10.1007/s00221-013-3419-2
21. Landy, M. S., Banks, M. S. & Knill, D. C. Ideal-observer models of cue integration. *Sensory Cue Integration Chapter 1* (Oxford University Press, New York, 2011). https://doi.org/10.1093/acprof:oso/9780195387247.001.0001
22. Buehner, M. J. Understanding the past, predicting the future: causation, not intentional action is the root of temporal binding. *Psychol. Sci.* **23**(12), 1490-1497 (2012). https://doi.org/10.1177/0956797612444612
23. Moore, J. W., Teufel, C., Subramaniam, N., Davis, G. & Fletcher, P.C. Attribution of intentional causation influences the perception of observed movements: behavioral evidence and neural correlates. *Front. Psychol.* **4**:23 (2013). https://doi.org/10.3389/fpsyg.2013.00023
24. Moore, J. & Haggard, P. Awareness of action: inference and prediction. *Conscious. Cogn.* **17**, 136-144 (2008). https://doi.org/10.1016/j.concog.2006.12.004
25. Synofzik, M., Vosgerau, G. & Voss, M. The experience of agency: an interplay between prediction and postdiction. *Front. Psychol.* **4**:127 (2013). https://doi.org/10.3389/fpsyg.2013.00127
26. Chambon, V., Sidarus, N. & Haggard, P. From action intentions to action effects: how does the sense of agency come about? *Front. Hum. Neurosci.* **8**:320 (2014). https://doi.org/10.3389/fnhum.2014.00320
27. Sidarus, N., Vuorre, M. & Haggard, P. Integrating prospective and retrospective cue to the sense of agency: a multi-study investigation. *Neurosci. Conscious.* **2017**:1 (2017). https://doi.org/10.1093/nc/nix012
28. von Holst, E. & Mittelstaedt, H. Das Reafferezprincip Wechselwirkungen zwischen Zentralnervensystem und Peripherie. *Naturwissenschaften* **37**, 464–476 (1950). [The reafference principle. In: *The Behavioral Physiology of Animals and Man. Selected Papers of E. von Holst*, Vol. 1. Martin, R.D., Trans. Coral Gables: Univ. of Miami Press].
29. Waszak F., Cardoso-Leite P. & Hughes G. Action effect anticipation: neurophysiological basis and functional consequences. *Neurosci. Biobehav. Rev.* **36**:943–959 (2012). https://doi.org/10.1016/j.neubiorev.2011.11.004
30. Haggard, P., Martin, F., Taylor-Clarke, M., Jeannerod, M. & Franck, N. Awareness of action in schizophrenia. *Neuroreport* **14**, 1081–1085 (2003). https://doi.org/10.1097/01.wnr.0000073684.00308.c0
31. Waters, F. & Jablensky, A. Time discrimination deficits in schizophrenia patients with first-rank (passivity) symptoms. *Psychiatry Res.* **167**(1-2), 12–20 (2009). https://doi.org/10.1016/j.psychres.2008.04.004
32. Voss, M., Moore, J., Hauser, M., Gallinat, J., Heinz, A. & Haggard, P. Altered awareness of action in schizophrenia: a specific deficit in predicting action consequences. *Brain* **133**, 3104–3112 (2010). https://doi.org/10.1093/brain/awq152