Supplementary material

Action-outcome learning and prediction shape the window of simultaneity of audiovisual outcomes

Andrea Desantis [1 *], Patrick Haggard [1]

[1]Institute of cognitive neuroscience, University College London, London, UK

Corresponding Author: *aerdna.desantis@gmail.com

Experiment 1

Participants. The sample size was a priori decided based on similar psychophysical experiments (i.e., 16 participants). Note that detecting audiovisual temporal asynchronies is challenging for some participants. Consequently, when participants in the first test session did not meet the performance level required to be included in the sample, they were not contacted to participate in the second session of the experiment. The criterion used to assess their performance was their temporal sensitivity to audiovisual asynchrony in the first session of the experiment. Based on this criterion, one participant was initially recruited but was not included in the sample. The subject did not complete the second session of the experiment due to very low temporal sensitivity to audiovisual asynchrony in the first session. Indeed, a high mean Standard Deviation calculated across all conditions was observed (mean SD was higher than the longest audiovisual SOA, i.e., SD > 266 ms). This indicates that the detection of audiovisual asynchrony was particularly challenging for that subject.

Data analyses

Fits were performed using *fitnlm,* a Matlab function fitting nonlinear regression models. We used a Gaussian model with 3 free parameters: 1) mean α, 2) standard deviation σ and 3) a scale factor *s,* which refers to the amplitude of the Gaussian curve.

$$f(x) = s \cdot e^{-(x-\alpha)^2/(2\sigma^2)}$$

Fitnlm implements a Least Square Estimation (LSE). Thus, in LSE, the sum of squares error (SSE) between observations and predictions is minimized.

The 10 audio-visual SOAs we used (-266, -133, -86, -66, -33, 33, 66, 86, 133, or 266 ms) were presented 12, 14, 16, 18, 20, 20, 18, 16, 14, and 12 times respectively (for a total of 160 trials).

Thus, in order to minimize the influence of errors on performances with the extreme SOAs, each *square error* was weighted according to the number of times that SOA was repeated.

Weights were determined by calculating the proportion of trials for each SOA (12, 14, 16, 18, 20, 20, 18, 16, 14, and 12) relative to the total number of trials (160). The same procedure was applied to all experiments.

Importantly, the function *fitnlm* in some cases estimated parameter *s* slightly higher than 1. This is unrealistic since *s* represents a probability (probability of judging the sound and the light as simultaneous). Thus, it cannot exceed 1. Consequently, in those cases we rerun our analyses constraining *s* parameter to not exceed 1. For this process we used Matlab function *fmincon*.

Results. Mean r^2 the four experimental conditions, as a measure of goodness-of-fit, is reported as follow: Action predicted pair M = 0.872; SD = 0.093; Action unpredicted pair M = 0.910, SD = 0.067; Sensory predicted pair M = 0.902, SD = 0.082; Sensory unpredicted M = 0.906, SD = 0.056.

D' values in the catch trials of the learning phase might not capture a difference between learning action-outcome associations and learning cue-audiovisual pair associations. However, an analysis of reaction times to detect unpredicted pairs in the learning phase might be more sensitive to highlight these differences. We performed a paired t-test on mean reaction times on catch trials of the learning phase of Experiment 1. Reaction times for correct detection of unpredicted pairs, when preceded by an action and when preceded by a sensory cue, were compared. The analysis showed that participants were overall faster in reporting an unpredicted audiovisual pair when this followed a visual cue (M = 552; SD = 36 ms) compared to when it followed an action (M = 523 ms; SD = 52 ms), F(1, 15) = 5.933, p = .028, η_p^2 = 0.283. This difference could be due to several factors. For instance, it could be due to the fact that the execution of a first key-press to trigger the audiovisual pair in the action condition might have slowed down participants when performing the double key-press to report the occurrence of an unpredicted pair.

Experiment 2

Results. Mean r^2 the four experimental conditions, as a measure of goodness-of-fit, is reported as follow: Action predicted pair M = 0.863; SD = 0.073; Action unpredicted pair M = 0.905,

SD = 0.066; Sensory predicted pair M = 0.895, SD = 0.065; Sensory unpredicted M = 0.887, SD = 0.068.

Experiment 3

Materials and Methods

Participants. Sixteen volunteers (11 women, average age = 22.93 years, *SD* = 2.74 years) participated in the experiment for an allowance of £ 7.5/h. All had normal or corrected-to-normal vision and hearing and were naïve as to the hypothesis under investigation. They all gave written informed consent. One participant was initially recruited but did not complete the full experiment.

Materials and Stimuli. See experiment 1

Procedure. The procedure was the same as for Experiment 1, except that action - audiovisual pair interval in the learning phase was fixed. Notably, in the learning phases the auditory and visual components of the audiovisual pair were always presented simultaneously, and with a delay of 330ms (i.e., mean interval calculated from the action-outcome intervals of the learning phase of Experiment 1) after the action or visual cue offset, respectively. The test phase was the same as the test phase of Experiment 1.

Results

A repeated measure ANOVA on PSS values with Action (present, absent) and Audiovisual pair (predicted, unpredicted) as factor was conducted. The analysis showed no significant interaction F(1, 15) = .091, p = .766, $\eta_p^2 =$ 0.006. Similarly, we did not observe any main effects of Action or Audiovisual pair, F(1, 15) = .087, p = .772, η_p^2 = 0.006, and F(1, 15) = .184, p = .674, $\eta_p^2 =$ 0.012, respectively.

The same analysis on temporal sensitivity (standard deviation of the psychometric function) values showed no interaction, F(1, 15) = .174, p = .682, $\eta_p^2 =$ 0.011, no main effect of Action F(1, 15) = 1.035, p =.325, η_p^2 = 0.064 and no main effect of Audiovisual pair F(1, 15) = 1.854, p = .193, $\eta_p^2 =$ 0.110.

Mean r^2 the four experimental conditions, as a measure of goodness-of-fit, is reported as follow: Action predicted pair M = 0.868; SD = 0.096; Action unpredicted pair M = 0.863, SD = 0.126; Sensory predicted pair M = 0.868, SD = 0.111; Sensory unpredicted M = 0.852, SD = 0.112.

We assessed whether in the learning phase participants learnt both action- and cue-audiovisual pair associations. A repeated measure ANOVA on identification *d'* for both action (*d': M* = 4.225, *SD* = 0.589) and sensory condition (*d': M* = 3.934, *SD* = 0.868) in the learning phase, showed no significant effect F(1, 15) = 3.92, p = .066, η_p^2 = 0.207. These results suggest that participants did learn both action and cue-stimulus associations, and that these associations were learnt with equal strength.

Finally, to assess whether participants were allocating the same amount of attentional resources to predicted/unpredicted audiovisual pairs in both the action and sensory conditions, we conducted a repeated measure ANOVA on identification performance in the catch trials of the test phase. The analysis showed no significant interaction F(1, 15) = .545, p = .472, η_p^2 = 0.035, no main effect of Action F(1, 15) = 2.893, p = .110, η_p^2 = 0.162, no main effect of Audiovisual pair F(1, 15) = .423, p = .525, η_p^2 = 0.027. This indicates that participants' attention was equally focused to stimuli in all conditions. Furthermore, the proportion of correct identification performances showed that stimuli were correctly identified in almost all catch trials (Action predicted pair: *M* = 0.962, *SD* = 0.041; Action unpredicted pair: *M* = 0.965, *SD* = 0.043; Sensory predicted pair: *M* = 0.957, *SD* = 0.040; Sensory unpredicted pair: *M* = 0.943, *SD* = 0.048).

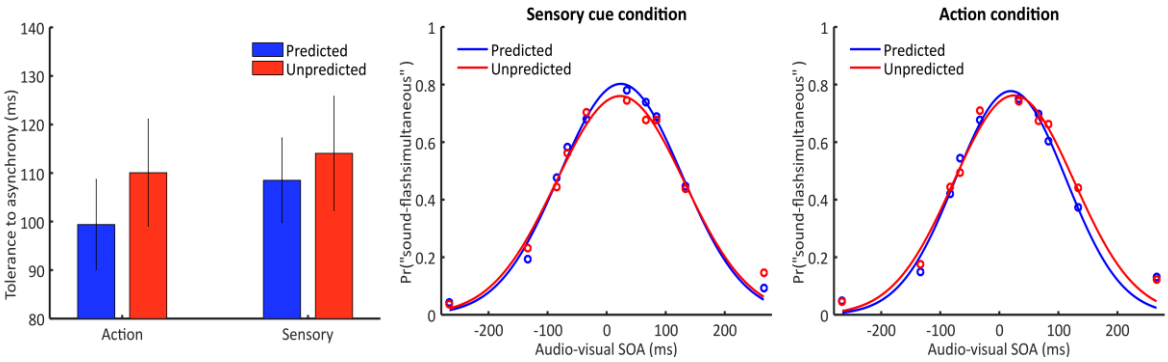

Figure 1. (Left panel) Mean tolerance to asynchrony (SD) values for all conditions (averaged across all participants). High SD values indicate high tolerance to audiovisual asynchronies, i.e., a wide WAS. (Central panel and right panel) Proportion of "sound and flash simultaneous" responses for predicted and unpredicted effects in the sensory cue and action condition, respectively (averaged across all participants) as a function of the 10 audiovisual SOAs.

Discussion

As for experiment 2 we did not observe any effects of prediction on temporal sensitivity values (measured by the standard deviation of the psychometric function): participants showed the

same sensitivity to audiovisual asynchrony in all conditions. Thus, having an accurate temporal prediction of the effect of an action cancelled the effect of the prediction of action-outcomes on temporal binding. This suggests that temporal prediction represents a strong cue indicating that two sensory outcomes should be integrated. Thus, the broad window of multisensory grouping found in Experiment 1 depends on predicting *what* will occur but not *when* it will occur. Acquiring a prediction of *when* outcomes would occur reduces the influence of action-outcome prediction on audio-visual binding.