# Mistaking opposition for autonomy: psychophysical studies on detecting choice bias

Ashild Kummen[1], Patrick Haggard[1], Gwydion Williams[1], Lucie Charles[1]

[1] Institute of Cognitive Neuroscience, University College London, Alexandra House, 17 Queen Square, London WC1N 3AZ, UK

Corresponding author:

Lucie Charles

Institute of Cognitive Neuroscience, University College London, Alexandra House, 17 Queen Square, London WC1N 3AZ, UK;

Email: l.charles@ucl.ac.uk

# ABSTRACT (200 WORDS LIMIT)

Do people know when they act freely and autonomously versus when their actions are influenced? While the human aspiration to freedom is widespread, little research has investigated how people perceive whether their choices are biased. Here, we explored how actions congruent or incongruent with suggestions are perceived as influenced or free. Across three experiments, participants saw directional stimuli cueing left or right manual responses. They were instructed to follow the cue's suggestion, oppose it, or ignore it entirely to make a 'free' choice. We found that we could bias participants' 'free responses' towards adherence or opposition, by making one instruction more frequent than the other. Strikingly, participants consistently reported feeling less influenced by cues to which they responded incongruently, even when response habits effectively biased them towards such opposition behaviour. This effect was so compelling that cues that were frequently presented with the *Oppose* instruction became systematically judged as having less influence on behaviour, artificially increasing the sense of freedom of choice. Taken together, these findings demonstrate that acting contrarian distorts the perception of autonomy. Crucially, we demonstrate the existence of a novel illusion of freedom evoked by trained opposition. Our results have important implications for understanding mechanisms of persuasion.

Keywords: Decision-making, Metacognition, Freedom of choice, Cognitive Control, Psychophysics

# 1 INTRODUCTION (621)

How much do we know about the true reasons for our choices, and about the causes of our decisions? Can we truly estimate if our choices are influenced by external factors?

Research in psychology has long demonstrated the limits of our voluntary control over our decisions: choices can be swayed by factors beyond our control, and beyond our awareness. For example, we may struggle to ignore sensory inputs even when we know these are irrelevant for the decision, and try to ignore them. This is true when visual attention is captured by distractors that present similarities with the target information like in the Eriksen Flanker Task, [1,2] and its variants [3,4] or when a target stimulus is preceded by a non-conscious prime, affecting response latency and choice [5].

Decisions are not only influenced by current information, but also by previous stimulation and previous responses. Learned stimulus-response associations influence future performance [6,7], and simple response repetition also biases current choices [8,9]. This influence of past behaviour continues even when it becomes irrelevant or unrewarding, and can only be overcome through additional cognitive control resources [10,11].This is illustrated in task switching paradigms in which participants frequently stick with the response they performed on the previous trial rather than use the response mapping required for the present trial [12].

While the mechanisms of unwanted influence on choice behaviour have been well documented [13] and formally modelled [14], it remains unclear whether people are able to detect when their choices are biased by such influences. Research in social psychology suggests that people are mostly unaware of the underlying drivers of their decisions [15,16]. Even when participants are conscious of using biased strategies in their decisions, feelings of objectivity prevail [17]. Equally, one might feel that extraneous influences are present even when they are not – as during "thought insertion" in psychosis [18]. In other words, introspective access to our own decision-making process may be limited.

Although "decision priming" has attracted wide scientific interest, few studies have investigated the subjective experience of external influence on decisions. When participants were shown ambiguous random dot kinematograms (RDK) and asked to respond 'freely', their actions were unsurprisingly influenced by small fluctuations in motion energy [19]. Strikingly, however, they reported stronger experiences of autonomy when they opposed the stimulus, than when they followed it [22].

In the present study, we have systematically investigated the link between opposition and subjective freedom. Across three experiments, we presented participants with random-dot kinematogram stimuli in which the majority of dots moved either leftward or rightward. Each display was accompanied by one of three instructions: respond congruently to the direction of the motion (Adhere condition), respond incongruently to it (Oppose condition), or make a free choice independent of the visual display (Detach condition). Thus, we could measure both the extent to which participants successfully ignored stimulus information (by quantifying the objective statistical relation between stimuli and response), and also their ability to introspect this stimulus-independence (by obtaining ratings of subjective freedom of choice).

We then used a structured series of experimental manipulations to investigate the cognitive basis underlying the sense of acting freely. In experiment 1, we presented participants more frequently with one instruction (Adhere or Oppose) than the other, to test if a learned response strategy would bias free choices. In experiment 2, we more specifically associated one direction of motion (left or right RDK stimuli) with one instruction (Adhere or Oppose) to test how using repeatedly one stimulus-response association would affect the perceived influence of that stimulus on behaviour. Finally in experiment 3, we increased the proportions of one direction of motion to test how the repetition of one visual cue would affect how participants perceived their freedom of choice when responding to it.

# 2 METHOD (2151)

## 2.1 EXPERIMENT 1

### 2.1.1 Participants

62 participants with normal or corrected-to-normal vision and no history of psychiatric of neurological illness were recruited. Sample size was based on a previous study [19] to detect an influence of signal fluctuations on free choice with 80% power at $\alpha = .05$ (two-tailed t-test for $d = 0.74$; required sample of at least 30 participants per group). The study was approved by the UCL Research Ethics Committee and participants were compensated £12. Two participants were excluded: Testing was interrupted for one participant as they failed to detect motion direction during training. One participant used one end of the freedom of choice scale in more than 98% of the trials and was therefore excluded from analysis. This left 60 participants (43 female; mean age 22.63 ± 3.17 years ranging from 20 to 30), randomly assigned to 30 in Adhere Group and 30 in Oppose Group.

### 2.1.2 Stimuli

Participants were presented with a series of random dot kinematograms (RDK) moving either to the left or to the right (see Stimuli in Supplementary Methods). The proportion of the dots moving in the same direction representing the motion coherence was adjusted by a Staircase Procedure to reach 71% coherence (see Staircase Procedure in Supplementary Methods). At a random time during stimulus presentation, the dots changed colour (red, green, or blue; matched for luminance and contrast) providing the instruction on how to respond in that particular trial. On (1) *Adhere* trials, participants were to respond congruently with the direction of the motion they perceived in the stimulus; on (2) *Oppose* trials, participants were to respond with the response key on the opposite side to the perceived motion direction; on (3) *Detach* trials, participants were to choose freely which response key to press, while trying to ensure that their choice was not influenced by the direction of dot motion. In those trials, participants were told not to use a set response strategy, such as responding always with the same key press or alternating between key presses, but try to act randomly.

### 2.1.3 Design & Procedure

In this experiment, one group of participants received the instruction to adhere (i.e. respond congruently to the motion direction) in most trials (Adhere group) while a second group received the instruction to oppose (i.e. respond incongruently) in most trials (Oppose group). This design allowed us to test whether a habitual response strategy would influence: (1) the objective ability to detach from the stimuli when given a free choice (objective FoC), and (2) the subjective feeling of acting freely (subjective FoC).

**Protocol**. The experiment followed a mixed design with group as the between-subjects variable (Adhere Group, Oppose Group), and instruction condition (Adhere, Oppose, Detach) as the within-subject variable. The main experiment consisted of 7 blocks of 57 trials, totalling to 378 trials of interest (excludes high coherence trials). For the majority Adhere Group, this was split so that 50% were Adhere trials, 16.7% Oppose, and 33.3% Detach. For the majority Oppose Group: 50% of trials were Oppose trials, 16.7% Adhere, and 33.3% Detach (see Figure 1). The experimental session lasted 80-90 minutes and was divided in two parts; the staircase procedure, and the main experiment.

**Trial procedure.** In the main experiment, participants started each trial with a fixation cross, before the stimulus was presented for 2500ms (see Figure 1A & B). At a random time-point between 100ms and 2388ms, the dots changed colour to indicate the instruction for this trial (see Figure 1C). Association between colours and instructions were counterbalanced across participants. After the stimulus disappeared, participants had 1500ms to respond with a left-arrow and right-arrow keypress using their right hand (see Figure 1C).

Following the response, participants were required to estimate how much their response choice was guided by the dot-motion, reporting their subjective experience of being more or less influenced by the stimulus (Figure 1E). To do so, a scale was presented for 3000ms and participants could move the slider with the '<' and '>' keys. The starting position of the slider was set at a random position on each trial. One end of the scale (counterbalanced across participants) indicated absolute independence of the stimulus ('I decided what to do myself, completely independently of what I saw on the screen"), while the other indicated complete dependent on the stimulus ("My response was determined entirely by what I saw on the screen"). Participants were instructed to use the entire range of the scale rather than only the extreme values and that their response should be guided by how they came up with their response on that given trials rather than the instruction.

### ***2.1.4*** Data Analysis

Trials in which no response was given were excluded (Adhere: $M$ = 4.20% of trials, $SD$ = 5.65%; Oppose: $M$ = 5.08%, $SD$ = 5.16%; Detach: $M$ = 6.60%, $SD$ = 6.88%). Each response was classified as congruent or incongruent with the stimulus direction (response mode factor). The mean proportion of motion congruent responses was calculated for each participant and each instruction. A repeated measure ANOVA was used to estimate the effects of group (between-participant factor) and instruction (within-participant factor) on the proportion of motion-congruent responses for instructed trials (i.e., Adhere and Oppose trials). A separate ANOVA was run for the Detach trials to estimate the effect of group (between-participant factor) on the proportion of motion congruent responses. Mean subjective FoC ratings were calculated for each combination of participant, instruction, and response mode. A repeated measure ANOVA was used to estimate the effect of group (between-participant factor), instruction and response mode (within-participant factor) on the reported freedom of choice in instructed trials. (Adhere and Oppose trials). Follow-up t-tests (two-tailed, unless specified otherwise) were used to compare ratings in error and correct trials for each instruction. A separate repeated measure ANOVA was run on the subjective FoC ratings in the Detach condition to estimate the effect of group (between-participant factor) and response mode (within-participant factor) on the reported freedom of choice.

## 2.2 EXPERIMENT 2

### 2.2.1 Participants

39 participants were recruited. Sample size was determined by a-priori power analysis based on the subjective FoC observed in experiment 3 (conducted prior to experiment 2), an effect size of d = 0.29 requiring N=37 to achieve 80% power, with alpha 0.05 (two-tailed t-test). One participant was excluded as they confused instructions. This left 38 participants (28 females; mean age 22.24 ± 2.94 years ranging from 18 to 30; all right-handed). Participants were randomly assigned to a group (20 for majority left motor responses;18 for majority right motor responses).

### 2.2.2 Design, Stimuli, & Procedure

Experiment 2 used a similar design than experiment 1 but that each stimulus direction was now associated more frequently with one instruction. Importantly, this also meant that participant used more frequently one motor response than the other (Figure 1). We therefore also included a second type of detach trial where the stimulus did not have any obvious motion direction (motion-neutral') to estimate the effect of purely motoric response repetition on the ability to detach.

The experiment followed a mixed design with group (majority left responses, majority right responses) as the between-subjects variable and instruction (Adhere, Oppose, Detach) as the within- subject variables. The stimuli and procedure for experiment 2 were identical to that of experiment 1, except for the modifications in the proportion of trials: In 469 trials, 59.7% of trials were instructed (half Adhere, half Oppose). Among instructed trials, the proportion of leftward and rightward stimuli differed between instructions. On Adhere trials, 75% of stimuli were in a direction congruent to the most frequent response of the group (e.g. rightward for the "majority right responses" group), whilst the remaining 25% were in the other direction. On Oppose trials, 75% of stimuli were in the incongruent direction to the most frequent response of the group (e.g. leftward for the "majority right responses" group), and 25% in the congruent direction (see Figure 1).

The remaining trials (40.3%) corresponded to the Detach instruction. Among the Detach trials, a third were moving leftward, a third moving rightward, and a third were motion-neutral where the motion coherence was set to 0% and *all* dots moved at random. These trials were added to investigate whether the repeated motor response would influence free choices in the absence of visual stimulation.

At the end of the experiment, participants were asked to estimate the overall proportion of right responses they had made (see Bias Awareness in Supplementary Results).

### 2.2.3 Data Analysis

Data analysis followed the same methods as for experiment 1. We separated trials according to whether the stimuli were frequently adhered, or frequently opposed. Trials in which no response were given were excluded (Adhere: *M* = 3.36%, *SD* = 3.14%; Oppose: *M* = 3.20%, *SD* = 3.70%; motion-present Detach: *M* = 5.33%, *SD* = 4.05%; motion-neutral Detach: *M* = 4.64%, *SD* = 3.84%). Mean proportions of motion congruent responses and mean subjective freedom of choice ratings were calculated per participant, split by instruction, stimulus-response association, and response mode and analysed using repeated measures ANOVAs (group as a between subject variable, and instruction, stimulus-response association, and response mode as within subject variables), similarly to Experiment 1. On motion-neutral detach trials, we tested whether small fluctuations in motion energy would influence response choice using reverse correlation (see Reverse Correlation Analysis in Supplementary Methods). Responses were classified by whether they were congruent or incongruent to the motion direction fluctuation and analysed following the same procedure as motion-present trials.

## 2.3 EXPERIMENT 3

### 2.3.1 Participants

42 participants were tested using similar procedures to experiment 1 and 2 (sample size indicated to achieve power of 80% for two- tailed t-test with effect size $d$ = 0.55 from experiment 1). Five participants were excluded. Two failed to see the dot motion direction during training, one used one end of the scale in more that 98% of trials or two had ceiling accuracy in instructed trials for rare stimulus direction, making it impossible to compare accuracy across response modes. This left 37 participants (24 females; mean age 24.35 ± 2.94 ranging from 18 to 30, all but five right-handed) allocated randomly to each group (20 participants: majority leftward stimuli; 17 majority rightward stimuli).

### 2.3.2 Design, Stimuli, & Procedure

To understand how repetition vs salience of stimuli may influence the ability to detach, the previous design was modified so that now one motion direction (leftward or rightward) was seen more frequently than the other. This allowed us to compare trials where a frequent stimulus was presented and trials where a rare stimulus was presented.

The experiment followed a mixed design with group (majority left-direction stimuli, majority right-direction stimuli) as the between-subjects factor, and instruction condition (Adhere, Oppose, Detach) and stimulus frequency (frequent stimulus, rare stimulus) as the within-subjects factors.

Stimuli and procedure were identical to those of experiment 1 and 2. The main modification was that 75% of all trials (except motion-neutral trials) had stimuli moving in the frequent direction (determined by participant's group), whilst the other 25% moved in the other direction (rare stimulus). A total of seven blocks with 60 trials of interest (excluding high coherence trials) was presented, giving a total of 420 trials of interest. In each block, 36 trials (60%) were instructed (half Adhere, half Oppose). The remaining 24 (40%) were detach trials; of which six (25%) were motion-neutral and 18 (75%) motion-present.

At the end of the experiment, participants were asked to estimate the proportion of leftward versus rightward stimuli (see Bias Awareness in Supplementary Results, for the findings).

### 2.3.3 Data Analysis

Trials with no response were excluded from the analysis (Adhere: $M$ = 8.34%, $SD$ = 10.38%; Oppose: $M$ = 8.71%, $SD$ = 10.91%; motion-present Detach: $M$ = 10.56%, $SD$ = 14.35%; motion-neutral Detach: $M$ = 10.94%, $SD$ = 16.45%). Mean proportion of motion congruent responses and mean subjective freedom of choice ratings were calculated per participant, split by instruction, stimulus frequency, and response mode and analysed using repeated measures ANOVAs (group as a between subject variable, and instruction, stimulus frequency, and response mode as within subject variables), similarly to Experiment 1&2. Motion-neutral detach trials were analysed following the same procedure as Experiment 2.

### 2.3.4 Data Analysis across experiments

In an exploratory analysis, we tested whether participants who more strongly used one response strategy in detach trial (for instance adhering more to the stimulus direction) would also report lower subjective freedom when using that same strategy in detach trials. To do so, we computed the difference in proportion of each participant's congruent and incongruent responses in the detach condition and tested whether a linear relationship was present with the difference in subjective freedom ratings between pro and incongruent in a regression model. To maximize power, we pooled the data across all experiments. We also performed the same regression in the specific parts of our design that most encouraged opposition during detachment (Oppose Group of experiment 1, frequently opposed stimuli condition of experiment 2, and frequent stimuli condition of experiment 3).

# 3 RESULTS

## 3.1 IMPACT OF DOMINANT INSTRUCTION ON FREEDOM OF CHOICE (EXPERIMENT 1)

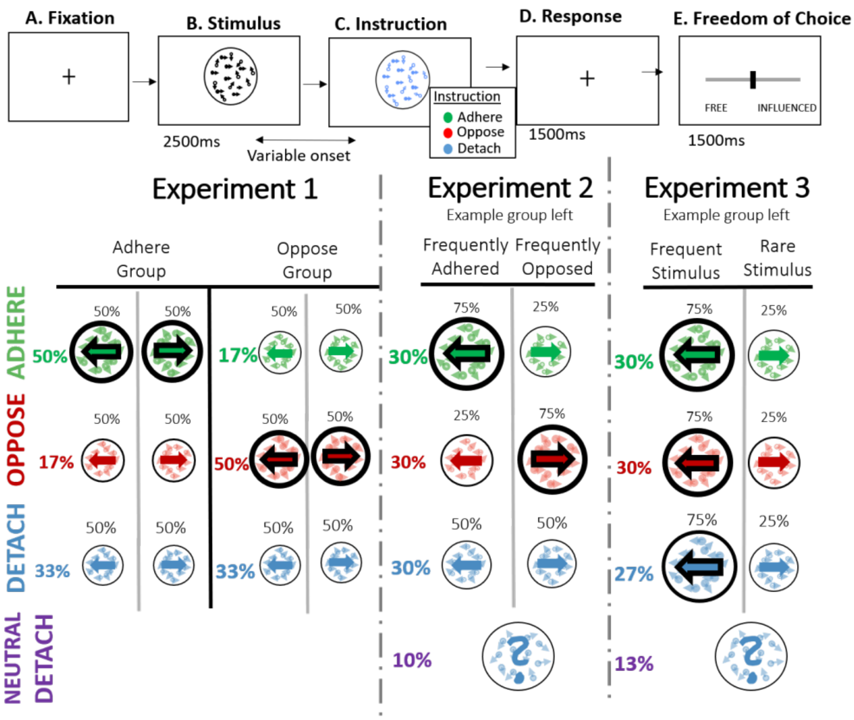


**Figure 1**. Experimental procedure (top) and design (bottom) of experiments 1-3. (Top) Following a fixation cross (A), the stimulus was presented for 2500ms (B). At a random timepoint during stimulation, the dots changed colour to indicate the instruction for that trial (Adhere, Oppose, or Detach; C). After giving their response (D), participants were asked to report their subjective feeling of freedom (E). (Bottom) Experiment 1 manipulated the frequency of instructions between groups so that one group were required to adhere more than to oppose (example highlighted by bold outline), while the other were required to oppose more than to adhere. Experiment 2 manipulated the total frequency of one response (example of Left motor-response group highlighted by bold-outline) by arranging for one stimulus direction to be preferentially associated with Adhere trials (frequently adhered) and the other stimulus direction with Oppose trials (frequently opposed). Experiment 3 manipulated the stimulus frequency so that each group saw more of one stimulus direction than the other (example of Left stimulus group highlighted by bold-outline).

Response accuracy was not affected by instruction frequency (Figure 2A) [*F*(1, 58) = 2.77, *p* = .102, $\eta_p^2$ = .046] or by group [*F*(1, 58) = 0.59, *p* = .447, $\eta_p^2$ = .010]. However, participants responded faster on Adhere than Oppose trials [*F*(1, 58) = 7.32, *p* = .009, $\eta_p^2$ = .112], as well as when they responded correctly compared to when they made an error *[F*(1, 58) = 66.25, *p*

< .001, $\eta_p^2$ = .533]. When they were instructed to detach from the stimulus (Figure 2B), participants in the Adhere group made more congruent responses on detach trials than participants in the Oppose group [$t(58) = 2.44$, $p = .018$, $d = 0.63$, two-tailed]. Considering each instruction separately, we observed non-significant trends for participants in the Oppose Group to make more incongruent responses ($M = 46.74\%$, $SD = 9.06\%$; $t(29) = -1.97$, $p = .06$, $d = -0.36$, two-tailed) while participants in the Adhere Group tended to respond congruently in 53.49% of detach trials ($M = 53.49\%$, $SD = 12.12\%$, $t(29) = 1.58$, $p = .13$, $d = 0.29$, two-tailed). Taken together, these results suggest that the dominant instruction affected detach trials and habitual response strategies continued to prevail even when attempting to detach.

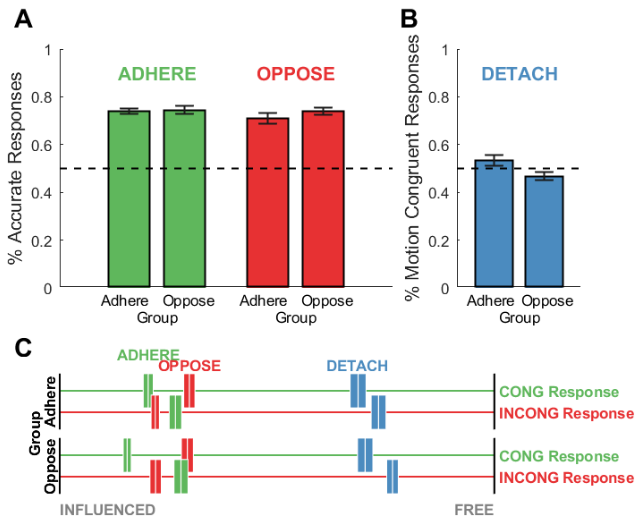


**Figure 2**. Results of Experiment 1. **A.** Mean proportion of correct responses on instructed trials (Adhere, green; Oppose, red). **B**. Proportion of motion congruent responses on Detach trials (blue) split by group (Adhere Group, Oppose Group). Error bars show standard error of the mean. **C**. Mean subjective freedom of choice ratings split by group (Adhere Group, Oppose Group), instruction (Adhere, Oppose, Detach), and response (congruent vs incongruent response; green and red lines respectively), for each stimulus motion direction. Ratings range from feeling completely influenced (left) to completely free (right). For instructed trials (Adhere: vertical green bars; Oppose: vertical red bars) matching bar and line colours (Adhere trials: green bars on green lines; Oppose trials: red bars on red lines) correspond to correct responses while bars on opposite colour lines (Adhere trials: green bars on red lines; Oppose trials: red bars on green lines) correspond to errors. Box width reflect +/- 1 standard error across participants..

We next considered how participants in each group rated their freedom of choice in instructed and free-choice trials. On instructed trials (Adhere and Oppose), participants reported feeling freer on trials where they were instructed to oppose the motion of the stimulus than when instructed to adhere to it [Figure 2C; main effect of instruction; $F(1, 58) = 11.43$, $p = .001$, $\eta_p^2$ = .165]. Moreover, they also reported feeling freer when making an error (e.g. incongruent responses for Adhere trials depicted by green bars on red lines on Figure 2C and congruent responses for Oppose trials depicted by red bars on green lines on Figure 2C) than when making an accurate response [interaction between instruction and response mode; $F(1, 58) = 88.42$, $p < .001$, $\eta_p^2 = .604$; Adhere instruction, Correct vs Error : $t(59) = -7.09$, $p < .001$, $d = -0.92$; Oppose instruction, Correct vs Error : $t(59) = -8.06$, $p < .001$, $d = -1.04$]. Participants may have paid less attention to the stimulus when they made an error than when responding correctly. An unseen or unattended stimulus cannot strongly influence behaviour. Thus, stronger sense of freedom suggests that participants used the subjective scale appropriately.

Turning to detach trials, we found that when making a response incongruent with the mean stimulus-motion direction, participants felt significantly freer (i.e., more detached from the stimulus) than when making a response congruent with the motion direction [Figure 2C; main effect of response mode; $F(1, 58) = 18.12$, $p < .001$, $\eta_p^2 = .238$]. Importantly, this was the case in the Adhere Group [$t(29) = 2.68$, $p = .012$, $d = 0.49$] but also in the Oppose Group [$t(29) = 3.32$, $p = .002$, $d = 0.61$], and the effect did not differ significantly between the groups [$F(1, 58) = 0.37$, $p = .54$]. Since the Oppose group made incongruent responses more frequently than the Adhere, this suggests that opposition was associated with increased sense of freedom, irrespective of whether opposition was the most frequent behaviour or not. In other words, acting contrarian was always associated with an increased sense of freedom, even when opposition was the dominant response tendency.

## 3.2 Impact of learned stimulus-response association on freedom of choice (Experiment 2)

While we showed that participants' objective ability to detach was partially compromised by following repeatedly one instruction over the other, we wondered whether this persistence effect would also be observed when one specific stimulus-response association was repeated, as opposed to one instruction. This was assessed in Experiment 2.

In instructed trials, we found that participants were more accurate in the condition corresponding to the trained stimulus-response association (Figure 3A). As in the first experiment, participants were also faster on Adhere than on Oppose trials [$F(1, 36) = 4.38$, $p = .044$, $\eta_p^2 = .108$], and faster for correct than incorrect responses [$F(1, 36) = 39.53$, $p < .001$, $\eta_p^2 = .523$]. When required to detach, (Figure 3B), we found that participants responded differently to stimuli that they had been trained to adhere and those they had been trained to oppose [$t(37) = 2.56$, $p = .015$, $d = 0.42$, two-tailed]: Participants tended to respond more congruently to frequently adhered stimuli [$M = 54.56\%$, $SD = 10.68\%$, $t(37) = 1.69$, $p = .050$, $d = 0.27$, one-tailed] and to respond more incongruently to stimuli they were trained to oppose [$M = 45.57\%$, $SD = 10.36\%$, $t(37) = 2.64$, $p = .006$, $d = 0.43$, one-tailed]. This result therefore replicated and extended those of Experiment 1, showing that training with one particular stimulus-response association continued to influence response strategy when required to detach. No effect of motor bias was observed in the motion-neutral trials [proportion of frequent responses compared to chance level on motion-neutral detach trials; $t(37) = 0.66$, $p = .510$, $d = 0.11$] and no further differences were observed when considering if the motion energy fluctuations favoured the frequently adhered or frequently opposed stimulus direction [Figure

3C; $t(37) = 0.65$, $p = .520$, $d = 0.10$], suggesting that purely motor repetition in itself did not affect the ability to detach.

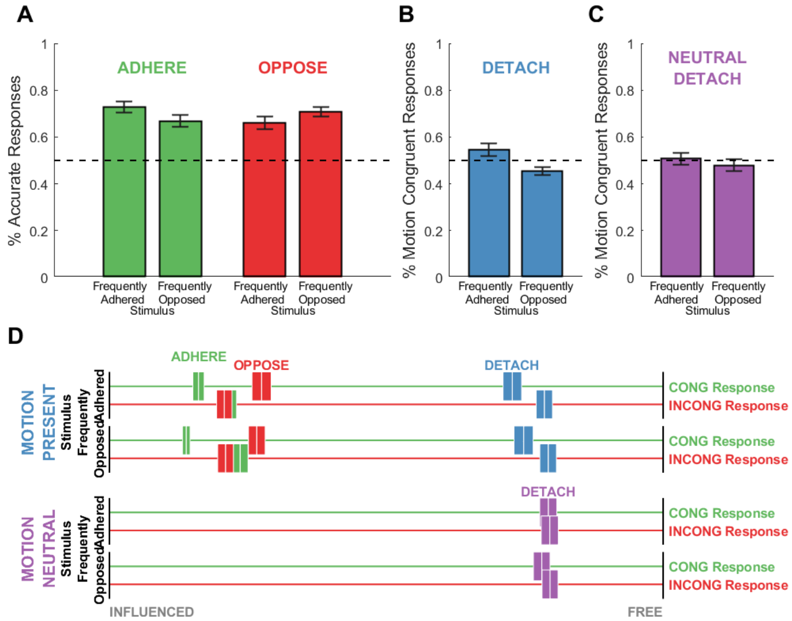


**Figure 3**. **Results of Experiment 2**. Accuracy for instructed trials (**A**) and percentage of motion congruent responses for motion-present Detach trials (**B**) and motion-neutral Detach trials (**C**) split by stimulus-response association (frequently adhered, frequently opposed). (**D**) Mean ratings of subjective ratings of freedom of choice split by stimulus-response association (frequently adhered, frequently opposed). Same legend as Figure 2.

Considering the ratings of freedom of choice (Figure 3D), we found that participants reported greater detachment on Oppose than on Adhere trials [$F(1, 36) = 5.13$, $p = .030$, $\eta_p^2 = .125$], and when making an error (e.g. incongruent responses for Adhere trials depicted by green bars on red lines on Figure 3C or a congruent response for Oppose trials depicted by red bars on green lines on Figure 3C; interaction between response mode and instruction: $F(1, 36) = 33.83$, $p < .001$, $\eta_p^2 = .484$; Adhere instruction, Correct vs Error $t(37) = -4.26$, $p < .001$, $d = -0.69$; Oppose instruction, Correct vs Error : $t(37) = -2.9$, $p = .006$, $d = -0.47$) replicating the results of our previous experiment. In detach trials (Figure 3D, blue bars), participants again reported they felt freer making a motion incongruent than a motion congruent response [$F(1, 36) = 12.55$, $p = 001$, $\eta_p^2 = .258$], confirming that opposition was always associated with an increased sense of freedom. Interestingly however, we also found an unexpected effect of previous opposition in detach trials: participants reported they felt more free when responding to a frequently-opposed stimulus compared to a frequently adhered stimulus [Main effect of stimulus-response association; $F(1, 36) = 4.34$, $p = .044$, $\eta_p^2 = .108$]. This was true regardless of the response made on the current trial [$F < 1$]. Hence, when attempting to detach, frequently-opposed stimuli became associated in themselves with an increased sense of autonomy, irrespective of whether participants actually opposed them on a given trial or not.

Interestingly, no effect of stimulus-response association or motor bias was observed on ratings of freedom in the motion-neutral detach trials (Figure 3D, purple, all *F* < 1) suggesting that motor novelty vs. motor repetition did not itself influence subjective freedom.

## 3.3 Impact of Stimulus Frequency on Freedom of Choice (Experiment 3)

We observed that participants reported feeling freer when presented with a stimulus they had frequently opposed in the past, compared to one frequently adhered to, regardless of their response to it. We then tested whether stimulus frequency itself impacted freedom of choice.

Stimulus frequency did not significantly affect response accuracy in instructed trials Figure 4A, although a trend was observed towards participants performing better for frequent stimuli [$F(1, 35) = 3.25$, $p = .08$, $\eta_p^2 = .085$]. As previously observed, participants responded more rapidly on Adhere trials compared to Oppose trials [$F(1, 35) = 6.88$, $p = .013$, $\eta_p^2 = .164$]. Accurate responses were again associated with faster response times than inaccurate responses [$F(1, 35) = 41.81$, $p < .001$, $\eta_p^2 = .544$].

In motion-present detach trials (Figure 4B, blue bars), no effect of stimulus frequency was observed on the proportion of motion congruent responses [frequent stimulus vs rare: $t(36) = 1.08$, $p = .290$, $d = 0.18$; frequent stimulus vs chance level: $t(36) = 1.18$, $p = .250$, $d = 0.19$; rare stimulus vs chance level: $t(36) = 0.48$, $p = .640$, $d = 0.08$], suggesting participants were equally likely to oppose dot-motion as they were to adhere to it, irrespective of whether the motion direction had been seen frequently or not. No further effects were observed on motion-neutral trials (Figure 4C), indicating participants had no response bias in the absence of a clear visual cue.

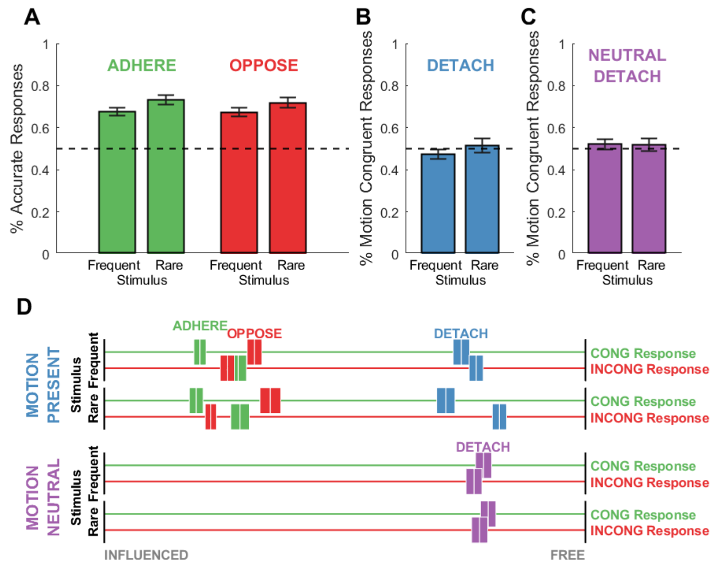


**Figure 4**. **Results of Experiment 3**. Accuracy for instructed trials (**A**) and percentage of motion congruent responses for motion-present Detach trials (**B**) and motion-neutral Detach

trials (**C**) split by stimulus frequency (frequent, rare). (**D**) Mean ratings of subjective ratings of freedom of choice split by by stimulus frequency (frequent, rare). Same legend as Figure 2.

As in our two previous experiments, we observed that freedom was rated higher on Oppose than Adhere trials [Figure 4D; $F(1, 35) = 10.29$, $p = .003$, $\eta_p^2 = .227$] and when making an error than when responding correctly [$F(1, 35) = 33.88$, $p < .001$, $\eta_p^2 = .492$; Adhere instruction, Correct vs Error $t(36) = -4.6$, $p < .001$, $d = -0.76$, $d = -0.69$; Oppose instruction, Correct vs Error : $t(36) = -3.6$, $p < .001$, $d = -0.59$]. Further, in detach trials, we replicated the findings of experiment 1 and 2 that participants reported feeling more free when making a response incongruent to the stimulus direction (Figure 4D, blue bars), compared to a congruent response [$F(1, 35) = 14.35$, $p < .001$, $\eta_p^2 = .291$]. Interestingly, we observed an interaction between stimulus frequency and response mode [$F(1,35) = 16.85$, $p < .001$, $\eta_p^2 = .325$], showing that participants reported feeling more free when making an incongruent response to a rare stimulus, compared to a frequent stimulus [$t(36) = 3.01$, $p = .005$, $d = 0.50$]. This suggested that the sense of freedom was increased when opposing a rare occurrence. No further effects were observed in motion-neutral detach trials.

## 3.4 Exploring correlations between subjective and objective freedom of choice

Finally, we tested whether participants who more strongly tended to adhere to the stimulus would also experience higher subjective freedom when opposing. Pooling data across all experiments, we found that this was indeed the case (Figure 5, regression slope = -19.62, beta = 5.31, $R^2 = .19$, $t(133) = 5.53$, $p < .001$; see Table S3 in Supplementary Materials). Interestingly, the intercept of this regression relation showed that, even when participants objectively detached, tending neither to oppose nor adhere, there remained a strong and significant tendency to rate incongruent responses as subjectively freer (beta = 5.32, $t(133) = 7.06$, $p < .001$, $d = 1.22$).

Performing the same regression in the specific parts of our design that most encouraged opposition during detachment (Oppose Group of experiment 1, frequently opposed stimuli condition of experiment 2, and frequent stimuli condition of experiment) however, revealed no significant correlation between the degree of detachment and the subjective sense of freedom for congruent versus incongruent responses (dataset with majority Oppose behaviour*:* $p = .130$). This suggests that this effect might be more strongly associated with adherence strategy. However, the intercept of the regressions remained significantly different from 0 (b = 5.41, $t(102) = 4.43$, $p < .001$, $d = 0.87$), suggesting that even when participants showed an objective tendency to oppose the stimulus, they still experienced greater freedom when opposing. Taken together, these results show that both habitual and contrarian behaviour modulated the sense of freedom of choice, independently of each other.

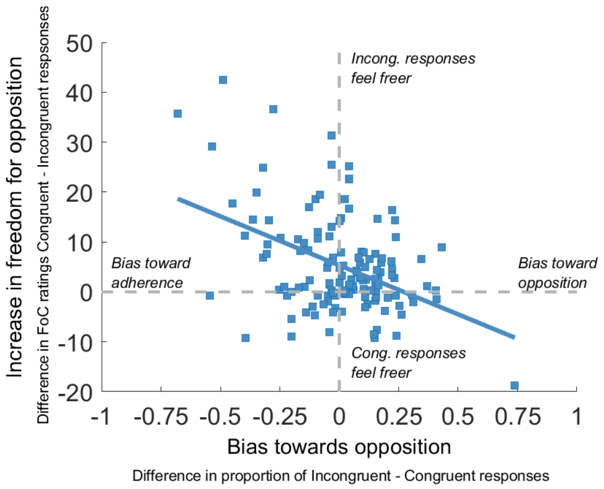


**Figure 5.** Difference in subjective freedom of choice (FoC) to incongruent minus congruent responses on motion-present detach trials according to the difference in proportion of incongruent minus congruent responses on motion-present detach trials. Each data-point represent one participant, pooling together all participants from all three experiments.

# 4 DISCUSSION (1358)

How do we know whether we are acting freely, or are influenced in our choices? And is our introspection of influence accurately reflecting our learned biases? In the present study, we tested how the tendency to respond congruently or incongruently to suggestion affected the sense of acting autonomously. To do so, we showed participants stimuli that primed a left or a right response and asked them to make a manual choice by either adhering to the direction suggested, opposing it, or detaching entirely from the visual stimulation to make a choice free from external influence. Crucially, we manipulated the stimulus and response distributions so that participants 1/ had to follow one set of instructions more frequently (experiment 1), 2/ had to respond to one stimulus with one response more frequently (experiment 2), or 3/ were exposed to one stimulus direction more frequently (experiment 3). This systematic set of designs allowed us to test how repeated association between a stimulus and a response could bias free decisions, and also how it could alter the subjective experience of acting freely.

Experiment 1 and 2 revealed that participants' ability to make free choices in the detach condition, was strongly compromised by implicitly learned stimulus-response associations. In Experiment 1, we observed that when participants were required to detach from the stimulus and attempted not to be influenced by it, they failed to do so. Instead, they tended to act in accordance with the instruction that they had been trained with the most frequently. This meant that stimulus-response mappings that were formed through repetition persisted in free choice. In experiment 2, participants were predominantly trained to make stimulus-congruent responses to some stimuli, while predominantly opposing others. Participants' free choices tended to perseverate in using the trained stimulus-response mappings. The ability to act autonomously and independently of a stimulus was therefore reduced if one had previously learned to make a particular response to that stimulus.

Such findings are in broad agreement with the classical concept of conditioned or habitual responding [21], as well as literature on task-switch costs. Only a few studies investigated how free choices themselves are impacted by previous experience of SR associations [22]. Our study shows that task history profoundly constrains the capacity for truly autonomous action. We show here that even when people try to ignore current stimulus content, their previous experiences strongly shape their behaviour. Importantly, persistence of SR mappings has also been found in subliminal priming paradigms [23], suggesting that it might operate independently of conscious processes of action selection [24]. Thus, the 'positive priming effect', in which participants tend to respond congruently to a subliminal prime [25–27] or be positively entrained by a non-conscious stimulus [28] could be reversed if the participants had been trained with the incongruent stimulus-response mapping [23]. Similarly, it has been observed that increasing the proportion of instructed trials strengthens the positive priming effect on free choices [29,30].

Our findings extend those results to free choices in the presence of supraliminal stimuli [31–34]. To act freely in the present study meant detaching from stimulus information. Participants however showed only a limited ability to dissociate from previously trained strategies. Crucially, we show that these are not merely effects of visual repetition alone, nor of motor priming alone. Thus, future free choices were influenced by the repeated association of one stimulus with a response, but not by patterns of stimulation alone, or by patterns of responding alone [35]. This finding suggests that volition and habit represent two dissociable, competing and mutually-exclusive routes to action [36] and that true detachment may be difficult to achieve once task habits are present, requiring effortful cognitive control to overcome habit-induced response conflicts [11]. Further research will be needed to understand the mechanisms of such effect and whether it is caused by occasional lapses in voluntary control, like the lapses of attention that cause errors in flanker tasks [11] or whether it corresponds to a more sustained "response priming" effect [23].

Importantly, our results bring new lights on whether participants are able to introspect the factors that influenced their choices. Crucially, we found that across all three experiments responding incongruently to the stimulus was associated with an increased subjective sense of acting freely and independently from the stimulus, compared to responses that followed the stimulus. Interestingly, this association between opposition and perceived autonomy was present irrespective of how much the stimulus in fact influenced participants' choices. In particular, whether participants more frequently followed the cued response, or more frequently opposed it, they always rated their choice as more autonomous when opposing the action suggested by the current circumstances. Indeed, even those participants who tended statistically to oppose the stimulus direction when required to detach nevertheless felt freer when they made incongruent responses on detach trials than when they made a congruent response (Experiment 1). This result was confirmed further by combining the conditions in each of the three experiments in which participants tended to make more opposing responses.

Inhibitory control therefore distorts the introspection of one's own choices: opposition increases the feeling of acting freely. A pooled regression further showed that this relation between opposition and subjective freedom was present even when controlling for participants' dominant response tendency, and for the time they took to make a choice. These findings replicate and extend previous results on the sense of autonomy and freedom [19]. Importantly, we demonstrate here that while free choices can be biased by experimental manipulation, the subjective bias caused by inhibitory control persists despite those manipulations. Thus, the subjective sense of autonomy seems to be driven largely by signals monitoring local conflict related to the external world, and processes that overcome those conflicts. Indeed, participants already rated being less influenced by the stimulus when they followed the Oppose instruction than when they were instructed to Adhere. The effect may recall the feeling of resisting external social influence, as when dissenting from the majority opinion [37] or disobeying norms such as traffic laws [43] [44,45]. The sense of thrill that accompanies rebellion or resistance to external influence, might be due to the metacognitive bias that makes acting contrarian feel like autonomy and freedom. As such, our findings might provide empirical evidence and potential cognitive mechanisms underlying social psychology phenomenon such as reactance [39,40]. Our results suggest that contrarian behaviours that aim to restore or boost subjective autonomy might actually stem from an underlying metacognitive bias in introspecting one's own decisions.

One interesting, and to our knowledge novel, result of this study concerns how participants linked an external stimulus to the feeling of acting freely. When participants learned by repeated experience to make stimulus-incongruent responses, they appeared to start associating that stimulus with an increased sense of freedom. Importantly, this effect was present irrespective of whether participants indeed acted independently from that stimulus or not on a specific trial (experiment 2). This finding shows that the sense of freedom that arose from opposing the influence of an external stimulus stuck to that stimulus, so that the stimulus itself becomes associated with acting freely. This result is striking because the experience of *not* being influenced was paradoxically evoked by the mere exposure to a visual cue. Such findings resonate with the literature on instrumental conditioning where a stimulus can become associated with a particular emotional [21] or cognitive response [41]. Our result could be interpreted as a novel illusion of autonomy, arising from people's limited ability to introspect the reasons for their choices [19]. This result raises the interesting possibility that the sense of autonomy can be increased artificially by training, since an illusory feeling of freedom could be induced by exposure to things that we have been told to reject. Such findings could shed new light on social effects such as irrational belief in conspiracy theories or extreme political radicalization. Indeed, a cognitive bias which conflates opposition and freedom could explain why people might feel autonomous when they are led into rebellions or manipulated into acting contrarian. More research will be needed to determine how our findings can be generalised to contexts of autonomous choice outside the laboratory, and how artificially generated feeling of freedom might influence behaviour.

# REFERENCES


1. Eriksen BA, Eriksen CW. 1974 Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychophysics* (doi:10.3758/BF03203267)

2. Stroop JR. 1935 Studies of interference in serial verbal reactions. *Journal of Experimental Psychology* (doi:10.1037/h0054651)

3. Williams JMG, Mathews A, MacLeod C. 1996 The Emotional Stroop Task and Psychopathology. *Psychological Bulletin* **122**, 3–24. (doi:10.1037/0033-2909.120.1.3)

4. Henik A, Tzelgov J. 1982 Is three greater than five: The relation between physical and semantic size in comparison tasks. *Memory & Cognition* **10**, 389–395. (doi:10.3758/BF03202431)

5. Schlaghecken F, Eimer M. 2004 Masked prime stimuli can bias quot;freequot; choices between response alternatives. *Psychonomic Bulletin and Review* **11**, 463–468. (doi:10.3758/BF03196596)

6. Talmi D, Seymour B, Dayan P, Dolan RJ. 2008 Human pavlovian-instrumental transfer. *Journal of Neuroscience* **28**, 360–368. (doi:10.1523/JNEUROSCI.4028-07.2008)

7. Rushworth MFS, Passingham RE, Nobre AC. 2002 Components of Switching Intentional Set. *Journal of Cognitive Neuroscience* **14**, 1139–1150. (doi:10.1162/089892902760807159)

8. Lages M, Boyle SC, Jaworska K. 2013 Flipping a coin in your head without monitoring outcomes? Comments on predicting free choices and a demo program. *Frontiers in Psychology* **4**, 925. (doi:10.3389/fpsyg.2013.00925)

9. Lages M, Jaworska K. 2012 How Predictable are “Spontaneous Decisions” and “Hidden Intentions”? Comparing Classification Results Based on Previous Responses with Multivariate Pattern Analysis of fMRI BOLD Signals. *Frontiers in Psychology* **3**, 56. (doi:10.3389/fpsyg.2012.00056)

10. Goschke T. 2003 Voluntary action and cognitive control from a cognitive neuroscience perspective. In *Voluntary action: Brains, minds, and sociality.*, pp. 49–85.

11. Botvinick M, Braver TS, Barch DDM, Carter CS, Cohen JD. 2001 Conflict monitoring and cognitive control. *Psychological Review* **108**, 624–652. (doi:10.1037/0033-295X.108.3.624)

12. Arrington CM, Logan GD. 2004 The cost of a voluntary task switch. *Psychological Science* **15**, 610–615. (doi:10.1111/j.0956-7976.2004.00728.x)

13. Newell BR, Shanks DR. 2014 Unconscious influences on decision making: A critical review. *Behavioral & Brain Sciences* **37**, 1–19. (doi:10.1017/S0140525X12003214)

14. Lavie N. 2010 Attention , Distraction , and Cognitive Control Under Load. (doi:10.1177/0963721410370295)

15. Pronin E, Lin DY, Ross L. 2002 The bias blind spot: Perceptions of bias in self versus others. *Personality and Social Psychology Bulletin* **28**, 369–381. (doi:10.1177/0146167202286008)

16. Epstein R, Robertson RE. 2015 The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections. *Proceedings of the National Academy of Sciences* **112**, E4512–E4521. (doi:10.1073/pnas.1419828112)

17. Hansen K, Gerbasi M, Todorov A, Kruse E, Pronin E. 2014 People Claim Objectivity After Knowingly Using Biased Strategies. *Personality and Social Psychology Bulletin* **40**, 691–699. (doi:10.1177/0146167214523476)

18. Frith CD. 2005 The neural basis of hallucinations and delusions. *Comptes Rendus - Biologies* **328**, 169–175. (doi:10.1016/j.crvi.2004.10.012)

19. Charles L, Haggard P. 2019 Feeling free: External influences on endogenous behaviour. *Quarterly Journal of Experimental Psychology* , 174702181988806. (doi:10.1177/1747021819888066)

20. Filevich E, Vanneste P, Brass M, Fias W, Haggard P, Kühn S. 2013 Brain correlates of subjective freedom of choice. *Consciousness and Cognition* **22**, 1271–1284. (doi:10.1016/j.concog.2013.08.011)

21. Maren S. 2001 NEUROBIOLOGY OF PAVLOVIAN FEAR CONDITIONING. *Annual Review of Neuroscience* **24**, 897–931. (doi:10.1146/annurev.psych.55.090902.141409)

22. Molina MT, Davis G. 2018 Unconscious priming dissociates ‘free choice’ from ‘spontaneous urge’ responses. *Consciousness and Cognition* **60**, 72–85. (doi:10.1016/j.concog.2018.02.003)

23. O’Connor PA, Neill WT. 2011 Does subliminal priming of free response choices depend on task set or automatic response activation? *Consciousness and Cognition* **20**, 280–287. (doi:10.1016/j.concog.2010.08.007)

24. Ocampo B. 2015 Unconscious manipulation of free choice by novel primes. *Consciousness and Cognition* **34**, 4–9. (doi:10.1016/j.concog.2015.03.007)

25. Le Bars S, Hsu YF, Waszak F. 2016 The impact of subliminal effect images in voluntary vs. stimulus-driven actions. *Cognition* **156**, 6–15. (doi:10.1016/j.cognition.2016.07.005)

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

27. Schlaghecken F, Eimer M. 2004 Masked prime stimuli can bias quot;freequot; choices between response alternatives. *Psychonomic Bulletin and Review* **11**, 463–468. (doi:10.3758/BF03196596)

28. Schurger A, Faivre N, Cammoun L, Trovó B, Blanke O. 2017 Entrainment of Voluntary Movement to Undetected Auditory Regularities. *Scientific Reports 2017 7:1* **7**, 1–7. (doi:10.1038/s41598-017-15126-w)

29. Bodner GE, Mulji R. 2010 Prime Proportion Affects Masked Priming of Fixed and Free-Choice Responses. *Experimental Psychology* **57**, 360–366. (doi:10.1027/1618-3169/a000043)

30. Fröber K, Dreisbach G. 2017 Keep flexible – Keep switching! The influence of forced task switching on voluntary task switching. *Cognition* **162**, 48–53. (doi:10.1016/j.cognition.2017.01.024)

31. Francken JC, Gaal S van, de Lange FP. 2011 Immediate and long-term priming effects are independent of prime awareness. *Consciousness and Cognition* **20**, 1793–1800. (doi:10.1016/j.concog.2011.04.005)

32. Kunde W, Reuss H, Kiesel A. 2012 Consciousness and cognitive control. *Advances in Cognitive Psychology*. **8**, 9–18. (doi:10.2478/v10053-008-0097-x)

33. Desender K, Van Lierde E, Van den Bussche E. 2013 Comparing Conscious and Unconscious Conflict Adaptation. *PLoS ONE* **8**, e55976. (doi:10.1371/journal.pone.0055976)

34. van Gaal S, de Lange FP, Cohen MX. 2012 The role of consciousness in cognitive control and decision making. *Frontiers in Human Neuroscience*. **6**, 121. (doi:10.3389/fnhum.2012.00121)

35. Maren S. 2001 Nuerobiology of Pavlovian fear conditioning. *Annual Review of Neuroscience*. **24**, 897–931. (doi:10.1146/annurev.neuro.24.1.897)

36. Haggard P. 2019 The Neurocognitive Bases of Human Volition. , 1–20.

37. Imhoff R, Erb H-P. 2009 What motivates nonconformity? Uniqueness seeking blocks majority influence. *Pers Soc Psychol Bull* **35**, 309–20. (doi:10.1177/0146167208328166)

38. Scott-Parker B, King MJ, Watson B. 2015 The psychosocial purpose of driving and its relationship with the risky driving behaviour of young novice drivers. *Transportation Research Part F: Traffic Psychology and Behaviour* **33**, 16–26. (doi:10.1016/j.trf.2015.06.004)

39. Steindl C, Jonas E, Sittenthaler S, Traut-Mattausch E, Greenberg J. 2015 Understanding psychological reactance: New developments and findings. *Zeitschrift fur Psychologie / Journal of Psychology* **223**, 205–214. (doi:10.1027/2151-2604/a000222)

40. Miron AM, Brehm JW. 2006 Reactance Theory - 40 Years Later. *Zeitschrift fur Sozialpsychologie* **37**, 9–18. (doi:10.1024/0044-3514.37.1.9)

41. Kirsch I, Lynn SJ, Vigorito M, Miller RR. 2004 The Role of Cognition in Classical and Operant Conditioning. *Journal of Clinical Psychology* **60**, 369–392. (doi:10.1002/jclp.10251)

42. Roitman JD, Shadlen MN. 2002 Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task. *The Journal of Neuroscience* **22**, 9475–9489. (doi:10.1016/S0377-2217(02)00363-6)

43. Kiani R, Hanks TD, Shadlen MN. 2008 Bounded Integration in Parietal Cortex Underlies Decisions Even When Viewing Duration Is Dictated by the Environment. *Journal of Neuroscience* **28**, 3017–3029. (doi:10.1523/JNEUROSCI.4761-07.2008)

44. Adelson EH, Bergen JR. 1985 Spatiotemporal energy models for the perception of motion. *Journal of the Optical Society of America A* **2**, 284. (doi:10.1364/JOSAA.2.000284)

## ACKNOWLEDGEMENT

We thank Ian Yeung for his help with the testing of participants.

## FUNDING

This work was supported by a British Academy Post-doctoral Fellowship to LC.


## AUTHORS CONTRIBUTIONS

A.K. contributed to designing the experiment, testing participants, analyzing data and writing the manuscript. P.H. and GW contributed to designing the experiment, interpreting data and writing the manuscript. L.C. contributed to designing the experiment, testing participants, analyzing data and writing the manuscript.

## COMPETING INTERESTS STATEMENT

The authors declare there is no competing interests.

Data is available at :

https://liveuclac-my.sharepoint.com/:f:/g/personal/ucjulch_ucl_ac_uk/EtMvbGTX8tFEqhgAtAPR-BcBx_vHiNPfuMsYr_DZoVB_Dg?e=XecnFr

Password: FoCContext

# 5 SUPPLEMENTARY MATERIALS

## *5.1* SUPPLEMENTARY METHOD

### *5.1.1* Stimuli

The dots were shown for a single frame before being replotted three frames later. When replotted, a subset of the dots were offset from their previous location to produce apparent coherent motion in one direction (to the right or to the left), whilst others where offset and plotted randomly (see [42]).

Each block of trials also included three high coherence trials (coherence set to 65%, one for each instruction type) to ensure that participants could follow instructions accurately in Adhere, Oppose, and Detach trials according to the colour change.

### 5.1.2 Staircase Procedure

To account for individual differences in sensitivity to dot motion, the motion coherence of the RDK stimuli was predetermined using a 2-down-1-up staircase procedure (Levitt', 1971). To do so, participants started the experiment with a block of 150 trials in which they were instructed to disregard all colour changes and to place their focus on detecting the direction of dot movement. The coherence level was adjusted starting at a motion coherence of 40% and after 12 initial trials, the motion coherence changed by a set amount in such a way that two consecutive correct guesses lowered the coherence, and a single error increased the coherence. The step size started at 6% and halved every time the participant made two consecutive errors (until the minimum step-size of 1.5%). The estimate of the appropriate motion coherence was obtained by averaging the coherence over the last 40 trials. This coherence level was used in the rest of the experiment.

### 5.1.3 Block Design

**Experiment 1.** The combination of orientation (left and right) and instruction type (Adhere, Oppose, Detach) generated 6 possible outcomes (i.e. orientation left to adhere; orientation right to adhere, etc) as an experimental trial in each group. Due to the between-subjects design, each group was assigned more Adhere trials than Oppose trials, or more Oppose trials than Adhere trials, whilst the proportion of Detach trials remained the same. Hence, Adhere Group had 50% Adhere trials and 17% Oppose trials (33% detach trials), whilst Oppose Group had 17% Adhere trials, 50% Oppose trials (and 33% Detach trials). Each block consisted of 54 trials consisting of the proportions mentioned above. This left a total of 378 trials of interest over 7 blocks. Additionally, each block of trials included 3 high coherence trials.

**Experiment 2.** The mixed experimental design was so that each group was assigned to either group majority left responses, or group majority right responses. Hence, the combination of orientation (left and right) and instruction type (Adhere, Oppose, Detach) was altered so that there was more Adhere for one direction, and more Oppose for the other direction. Using group left as an example, the Adhere trials (30% of all trials) consisted of 75% left-direction stimuli, and 25% right-direction stimuli. On Oppose trials (30% of all trials), 75% was right-direction stimuli and 25% left-direction stimuli. The percentage of Detach trials was 40%, of which 75% were motion-present at 25% motion-neutral. For each group, there was a total of 67 randomly arranged experimental trials of interest per block (and a total of 469 trials of interest over 7 blocks) following the proportions mentioned above. Additionally, each block of trials included 3 high coherence trials.

**Experiment 3.** Again, the mixed experimental design was so that participants were assigned to see either left-ward stimulus frequently, or right-ward. Hence, the combination of orientation (left, right) and instruction type (Adhere, Oppose, Detach) was so that there was more of one direction than the other across all three instruction types. Overall, 75% of all trials was one direction (e.g. leftward for group majority left), and 25% the other (e.g. rightward for group majority right. 30% of trials were Adhere, 30% were Oppose, 27% Detach (two thirds motion-present at one third motion-neutral). For each group, there was a total of 60 randomly arranged experimental trials of interest per block (and a total of 420 trials of interest over 7 blocks) following the proportions mentioned above. Additionally, each block of trials included 3 high coherence trials.

### 5.1.4 Reverse correlation analysis on motion-neutral detach trials

On motion-neutral detach trials, small fluctuations in motion energy could still momentarily favour one direction over the other, even though overall motion energy was balanced across directions. The time-course of the motion energy for each trial was retrieved by applying spatiotemporal motion filters to dot position over time so to compute a time-course of motion direction for each trial [19,43,44]. The net motion energy value was calculated by subtracting the amount of rightward motion from leftward motion on each timeframe of stimulus presentation. These values were then standardized against the mean motion energy and standard deviation of all other motion-neutral detach trials to give a zero mean and unit of standard deviation. Thus, motion energy values now reflected fluctuations in motion direction around the mean coherence level. We then reverse correlated the normalized values so that positive values related to fluctuations favouring the frequently adhered direction (Experiment 2) or the frequent direction (Experiment 3), whilst negative values reflected fluctuations favouring the frequently opposed direction (Experiment 2) or the rare direction (Experiment 3). The sum across time (area under curve, AUC) was computed. We then divided trials according to whether responses were classified by whether they were congruent or incongruent to the direction of this fluctuation.

## 5.2 Supplementary Results

### 5.2.1 Bias awareness

**Experiment 2.** All participants were asked to estimate the proportion of right responses across all trials of the experiment on a scale from 0 (only left responses) to 100 (only right responses). An estimate of 50 would reflect 50/50 left and right responses. A one-sample t-test was performed for each group, comparing the estimated perceptual bias to chance level (50%). Both descriptive statistics and t-test comparisons are reported in Table S1. The null effect for both groups meant that participants did not feel they responded more of one button than the other, and hence were not aware of our response bias manipulation.

Table S1

Bias Awareness: Means and Standard Deviation for Estimated Proportion of Right Responses (by group), and T-test Comparison to Chance Level (50%)

| Group | M | SD | t | df | p | d |
|---|---|---|---|---|---|---|
| Group Majority Left Responses | 48.80% | 17.21% | 0.31 | 19 | .760 | 0.07 |
| Group Majority Right Responses | 52.33% | 19.39% | 0.51 | 17 | .620 | 0.12 |

**Experiment 3.** All participants were asked to estimate the proportion of right-direction trials on a scale from 0 (all left-direction trials) to 100 (all right-direction trials). A one-sample t-test was performed for each group, comparing the estimated perceptual bias to chance level (50%). Both descriptive statistics and t-test comparisons are reported in Table S2. The null effect for group majority leftward stimuli suggest that participants did not feel they saw either direction more than the other, and hence were not aware of our stimulus frequency manipulation. However, group majority rightward stimuli reported significantly more right-direction trials than left-direction trials, suggesting an awareness of the stimulus frequency manipulation

Table S2

Bias Awareness: Means and Standard Deviation for Estimated Proportion of Right-direction Trials (by group), and T-test Comparison to Chance Level (50%)

| Group | M | SD | t | df | p | d |
|---|---|---|---|---|---|---|
| Group Majority Leftward Stimuli | 48.70% | 19.99% | 0.291 | 19 | .770 | 0.07 |
| Group Majority Rightward Stimuli | 59.29% | 13.95% | 2.75 | 16 | .014 | 0.67 |

### 5.2.2 Reaction Time on Detach Trials

**Experiment 1.** There was no main effect of response mode for reaction time on detach trials responses [F(1, 58) = 2.21, p = .142, $\eta_p^2$ = .000], reflecting that a congruent response was no easier made than an incongruent response. There was also no main effect of group [F(1, 58) = 0.01, p = .906, $\eta_p^2$ = .000], or an interaction between group and response mode [F(1, 58) = 0.12, p = .730], suggesting that being trained for one strategy (Adhere Group vs Oppose Group) does not influence the ease at which participants go with or against that strategy on detach trials as measured in reaction time.

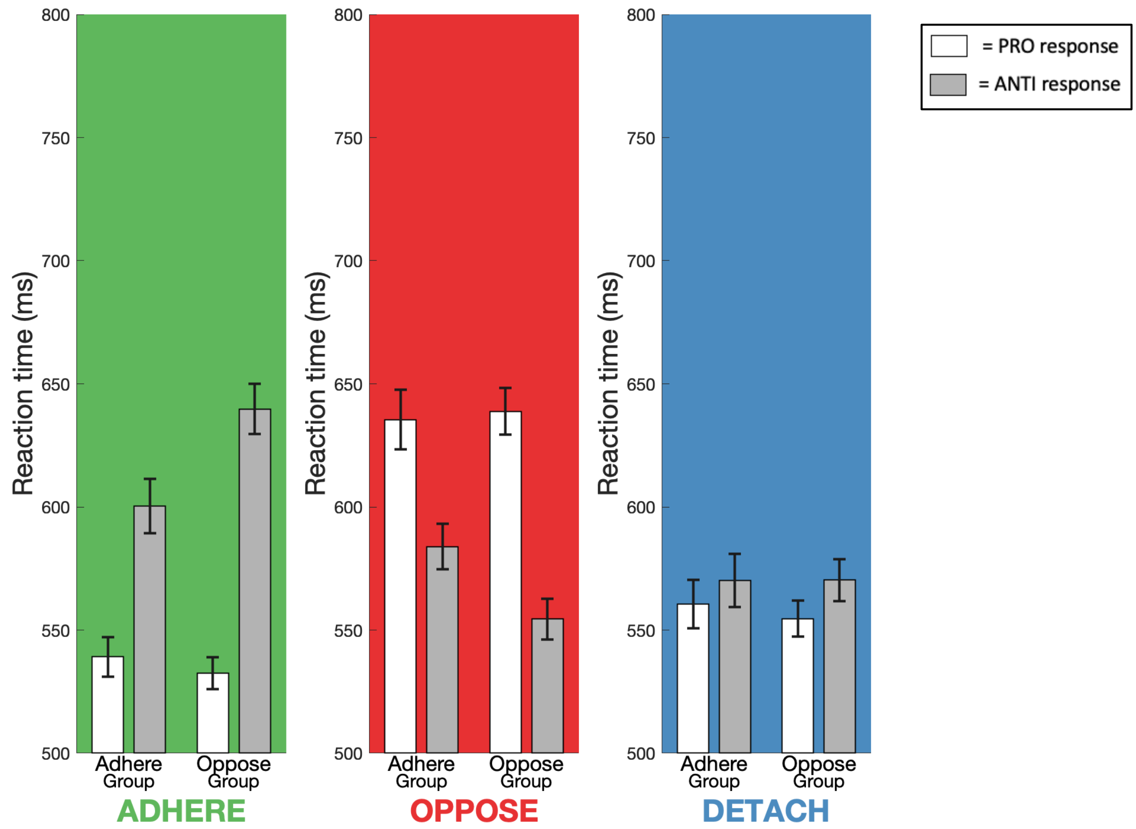


**Figure S1**. Mean reaction time in milliseconds on Experiment 1 split by instruction (Adhere, Oppose, Detach), group (Adhere Group, Oppose Group), and response mode (congruent responses, white bars; incongruent responses, grey bars).

**Experiment 2.** On motion-present detach trials, there was no difference in the reaction time between a congruent and an incongruent response [main effect response mode; $F(1, 36) = 1.70$, $p = .201$, $\eta_p^2 = .045$]. There was also no main effect of stimulus-response association [$F(1, 36) = 0.04$, $p = .846$, $\eta_p^2 = .001$], or an interaction between stimulus-response association and response mode [$F(1, 36) = 1.28$, $p = .266$, $\eta_p^2 = .034$]. This suggests that being trained for one strategy at one particular stimulus does not influence the ease at which participants go with or against that strategy on detach trials as measured in reaction time.

The results on motion-neutral detach trials were similar to that of motion-present detach trials, with no main effect of response mode [$F(1, 36) = 0.73$, $p = .398$, $\eta_p^2 = .019$] stimulus-response association [$F(1, 36) = 0.23$, $p = .634$, $\eta_p^2 = .006$], or interaction between response mode and stimulus-response association [$F(1, 36) = 2.50$, $p = .122$, $\eta_p^2 = .063$]. This suggests that regardless of fluctuations in motion energy favouring one direction over the other, or whether a motor response is one made frequently or rarely over past trials, there is no difference in the ease at which participants make such a response on detach trials.

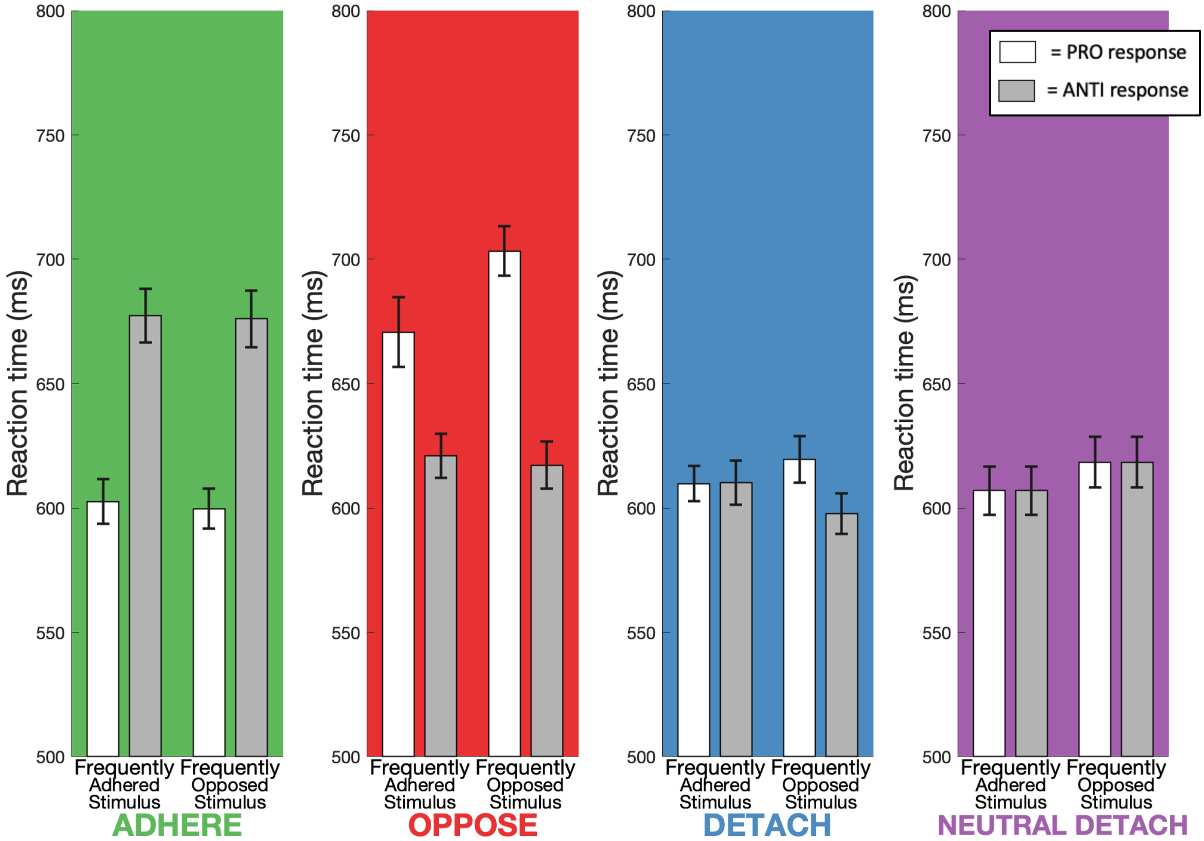


**Figure S2**. Mean reaction time in milliseconds on Experiment 2 split by instruction (Adhere, Oppose, Detach), stimulus-response association (frequently adhered, frequently opposed), and response mode (congruent responses, white bars; incongruent responses, grey bars).

**Experiment 3.** Again, there was no difference in the reaction time on detach trials between a congruent and an incongruent response [main effect response mode; $F(1, 35) = 0.39$, $p = .538$, $\eta_p^2 = .011$]. There was also no main effect of stimulus frequency [$F(1, 35) = 0.81$, $p = .373$, $\eta_p^2 = .023$]. This suggests that seeing one stimulus direction more often than the other does not influence the ease at which one can detach from that stimulus as reflected in reaction times. There was, however, a 3-way interaction with group, response mode, and stimulus frequency, which is reported and discussed in Supplementary Results.

Motion-neutral detach trials also had no difference in reaction time between response modes [$F(1, 35) = 0.07$, $p = .788$, $\eta_p^2 = .002$] or stimulus frequency [$F(1, 36) = 0.16$, $p = .691$, $\eta_p^2 = .005$]. Hence, all detach trials suggested that the frequency at which a cue is presented does not influence the ability to detach from that cue as measured in reaction time.

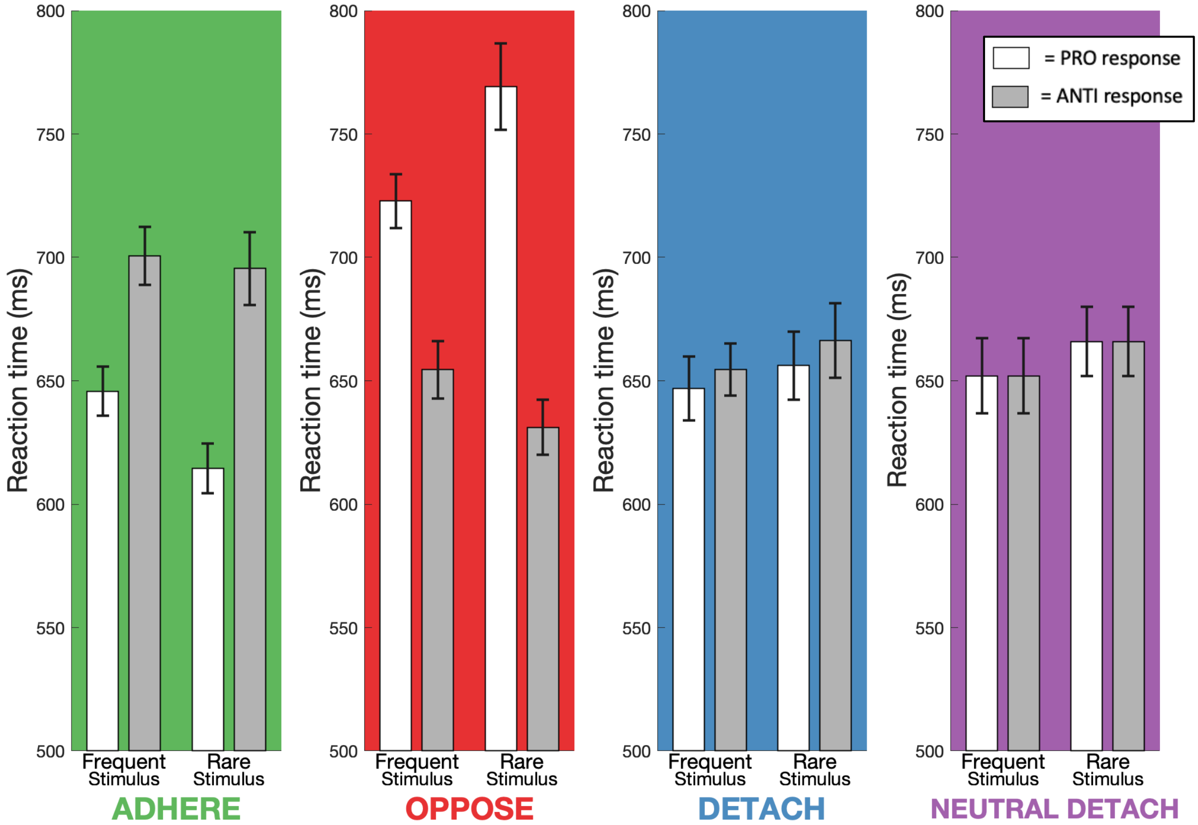


**Figure S3**. Mean reaction time in milliseconds split by instruction (Adhere, Oppose, Detach), stimulus frequency (frequent, rare), and response mode (congruent responses, white bars; incongruent responses, grey bars).

## 5.3 Group Differences on Experiment 2 and 3

**Experiment 2.** The same was true for experiment 2, where participants were assigned to press either more of the right button than the left, or more of the left button than the right. Whether making more of a right or left button press influenced the ability to detach or subjective rating of detachment was not of interest to our hypotheses. We did however observe group-dependent findings for accuracy on instructed trials [interaction group and stimulus-response association: $F(1, 36) = 7.91$, $p = .008$], as well as proportion of motion congruent responses on detach trials [interaction group and motion coherence: $F(1, 36) = 4.16$, $p = .049$].

The group differences observed were perhaps due to a population-based preference for making one motor response over the other, which interacted with our response frequency manipulations.

**Experiment 3.** For experiment 3, participants were assigned to either see mostly leftward stimuli, or mostly rightward stimuli. We were interested in the ability to detach and subjective introspection of detachment from the stimulus, and how it differed between a frequent and a rare stimulus. Group differences in our statistical analyses, reflecting that seeing a lot of right-ward motion compared to left-ward motion influenced subjective and objective FoC, were not of interest in regard to our hypotheses. However, we observed group difference in on instructed trials for subjective rating of freedom [interaction group, instruction, and stimulus frequency: $F(1, 36) = 5.03$, $p = .031$], as well as accuracy [interaction group and stimulus frequency: $F(1, 36) = 7.29$, $p = .011$].

The group differences observed were likely due to increased or decreased sensitivity to one motion over the other. Previous research has indicated that subjects that have English or another left-to right reading language have a higher sensitivity for leftward motion (Morikawa & McBeath, 1991), but more research is needed to determine the cause of our group effects.

## 5.4 Regression

**Table S3**

Regressing difference in subjective freedom of choice from incongruent to pro responses with motion consistency and difference in reaction time (RT) as predictors. Datapoints are combined across experiments.

| | b | t | df | p |
|---|---|---|---|---|
| Adhere Group, frequently adhered stimuli, rare stimuli combined | | | | |
| Intercept | 6.57 | 6.17 | 102 | .000 |
| Motion Consistency* | -9.87 | 2.73 | 102 | .008 |
| RT difference pro | 0.02 | 2.25 | 102 | .026 |
| Oppose Group, frequently opposed stimuli, frequent stimuli combined | | | | |
| Intercept | 5.41 | 4.28 | 102 | .000 |
| Motion Consistency* | -13.36 | 2.54 | 102 | .130 |
| RT difference incong-cong | -0.01 | 0.81 | 102 | .420 |

*Note*. Motion consistency measured as proportion of incongruent responses minus the proportion of congruent responses on detach trials.

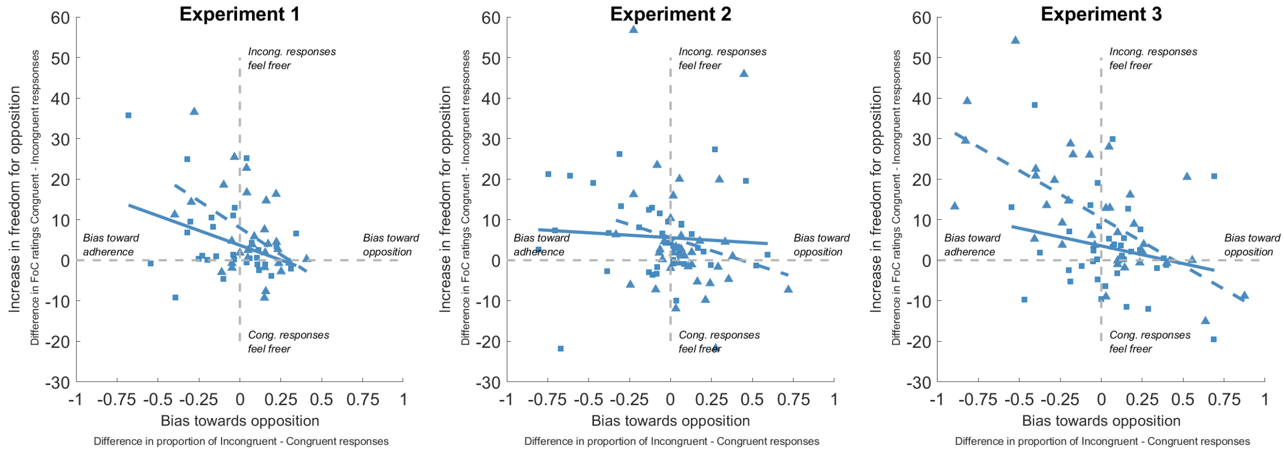


**Figure S4**. Difference in subjective freedom of choice (FoC) to incongruent minus congruent responses on motion-present detach trials correlated with difference in proportion of incongruent minus congruent responses on motion-present detach trials. This is split by each experiment, and the experiment-specific main manipulation (Adhere Group vs Oppose Group on Experiment 1; frequently adhered vs frequently opposed stimuli on Experiment 2; frequent vs rare stimuli on Experiment 3)

## 5.5 Supplementary References

Morikawa, K., & McBeath, M. K. (1992). Lateral motion bias associated with reading direction. *Vision Research*, *32*(6), 1137–1141. https://doi.org/10.1016/0042- 6989(92)90014-A

Levitt, H. (1971). Transformed Up-Down Methods in Psychoacoustics. *The Journal of the Acoustical Society of America*, *49*(2B), 467–477. https://doi.org/10.1121/1.1912375