# The Role of Conjunctive Representations in Controlling Actions

Atsushi Kikumoto[1] and Ulrich Mayr[2]

[1] Brown University, [2] University of Oregon

Abbreviated Title: Conjunctive Representations and Action Control

Corresponding Author: Ulrich Mayr
E-mail: mayr@uoregon.edu

We declare no financial interests or conflicts of interest.

Keywords: conjunctive representations, stop-signal task, EEG

This research was supported by NIA grant R01 AG037564-01A1, and by NSF grant 1734264.

**Abstract**

Prominent theories of action control propose that conjunctive representations, which integrate task-relevant features in a nonlinear manner, are critical for successful action control. Thus, in order to stop an initiated action, which is a key aspect of self-control, conjunctive representations should be the primary target of the stopping process. We tested this hypothesis by combining a rule-based action selection task with the stop-signal paradigm. Participants selected actions based on abstract stimulus-response rules and occasionally received a stop-signal as a prompt to halt the intended action. Using time-resolved representational similarity analysis of the EEG signal, we decoded both orthogonal, constituent action-relevant representations (rules, stimuli, and responses) and conjunctions of these features in a time-resolved manner and on the level of single trials. In Exp. 1, where a short stop-signal interval (100 ms) ensured high stopping success, simple action-relevant features and their nonlinear conjunction were robustly expressed in the EEG signal. Importantly, the conjunctive representation was selectively suppressed on stop trials. In Exp. 2, an adaptive staircase targeting a stopping success of 50%. Here, conjunctions were selectively suppressed on successful stop trials compared to both go and failed stop trials. Moreover, the strength of conjunctive representations at the time of the stop-signal uniquely predicted stopping failures. Combined, these results clarify that the stopping process does not just target motor output representations. Rather, conjunctive representations seem to be critical for selecting a specific action and for that reason also need to be suppressed by the stopping process in order to cancel intended actions.

## Significance Statement

The neural-cognitive machinery behind stopping an intended action—a critical aspect of self-control—is relatively well understood. However, we do not know the exact target of this stopping processes: Does it simply suppress motor output? Or does it affect higher-level representations? Using EEG decoding techniques, we show that stopping affects a highly integrated, conjunctive representation that nonlinearly combines all action relevant features (stimuli, responses, and rules). Further, the stronger this representation is expressed, the harder it is to strop the intended action. These results are consistent with the hypothesis that conjunctive representations are a necessary precursor for successfully executing a specific action. For that reason, suppressing these representations is the key to stopping an initiated action.

## The Role of Conjunctive Representations in Controlling Actions

Even simple goal-directed actions, such as kicking a soccer ball to a team-mate, rely on various sensorimotor features—the location of the ball, the presence of opponent players and teammates, as well as on abstract rules (e.g., “kick softly when the grass is wet”). In traditional, stage-type information processing models, such different aspects are handled independently, and in a serial, feed-forward manner (Posner and Mitchell, 1967); Donders (1969); (Sternberg, 1969; Kornblum et al., 1990; Sanders and Sanders, 2013). Alternatively, there are also models in which all action-relevant features are combined within a common representational space during selection. Specifically, event-file theory (Hommel, 1998; Hommel et al., 2001; Schumacher and Hazeltine, 2016; Hommel, 2019) posits that an action becomes executable only once all task-relevant features are integrated within conjunctive representations (i.e., event files). Moreover, recent research in non-human primates indicate that neurons with nonlinear, mixed selectivity response properties integrate various aspects in a conjunctive manner and play a causal role in successful action (Rigotti et al., 2013; Parthasarathy et al., 2017).

If conjunctive representations are necessary, and maybe even sufficient precursors of goal-directed behavior, it follows that the pathway towards controlling a given action needs to lead through these representations. For example, when in the earlier example, an opponent defender suddenly blocks the targeted teammate, the act of kicking needs to be canceled. The cognitive and neural underpinnings of action inhibition have been well characterized using variants of the stop-signal paradigm (Logan and Cowan, 1984; Swann et al., 2009; Aron et al., 2014; Verbruggen et al., 2019b; Wessel, 2019). Yet, it is currently an open question how stopping affects the different representations that underlie planned actions. In theory, the stopping process might occur simply by suppressing response-related representations that directly link to motor control pathways (Coxon et al., 2006; Labruna et al., 2014; Greenhouse et al., 2015; Duque et al., 2017), while leaving other, higher-level representations intact.

However, assuming conjunctive representations are indeed critical for action control, the cancellation of an initiated action should require the suppression of the entire, integrated representation of the action plan.

Kikumoto and Mayr (2020) recently applied a time-resolved representational similarity analysis (RSA) (Kriegeskorte et al., 2008) to the EEG signal in order to track the presence of conjunctive and constituent feature representations during rule-based action selection. These analyses indeed revealed conjunctive representations that integrated action rules to specific sensory/motor settings throughout the entire selection period. Moreover, conjunction strength was a robust and unique predictor of trial-to-trial variability in RTs—as one would expect if conjunctive representations are necessary and sufficient conditions for action execution.

To directly test the hypothesis that cancelling conjunctive representations is the pathway to action inhibition, in the current work, we combined a rule-based action selection task (Fig. 1ab) (Mayr and Bryck, 2005; Kikumoto and Mayr, 2020) with an occasional stop-signal (Verbruggen et al., 2019a). As in Kikumoto and Mayr (2020), we used RSA to track action-relevant representations. In Exp. 1, the stop-signal was presented 100 ms after the stimulus onset, which is early enough for successful stopping on most trials. In Exp. 2, stop-signal timing was adjusted via a staircase-tracking procedure (Verbruggen et al., 2019a), based on participants' trial-to-trial stopping accuracy . Across both experiments, we found strong evidence that successful stopping selectively suppresses conjunctive representations. In Exp. 2, we also found that the strength of conjunctive representations at the time the stop-signal arrives, is a unique predictor of stopping success.

## Materials and Methods

### *Participants*

A total of 64 people participated after signing informed consent following a protocol approved by the University of Oregon's Human Subjects Committee in exchange for the compensation of $10 per hour and additional performance-based incentives. Participants with

excessive amount of EEG artifacts (i.e, more than 35% of trials; see EEG recordings and preprocessing for detail) were removed from further analysis. As a result, we retained 24 out of 26 participants for Exp. 1 and 36 out of 38 for Exp. 2. In Exp. 2, 3 participants were further excluded because of failures to stop in excess of 75% (Verbruggen et al., 2019a).

***Stimuli, Tasks and Procedure***

The task combined a variant of rule-based action task (Mayr and Bryck, 2005) with the stop-signal paradigm (Verbruggen et al., 2019a). Participants were randomly cued on a trial-by-trial basis to execute one of the three possible actions rules (Fig. 1a). Based on the cued rule, participants responded to the location of a circle (1.32° in radius) that randomly appeared in the corner of a white frame (6.6° in one side) by selecting one of the four response keys that were arranged in 2 x 2 matrix (4, 5, 1, and 2 on the number pad). For example, the vertical rule mapped the left-top dot to the bottom-left response as a correct response. Two different cue words (e.g., vertical and updown) were used for each rule (i.e., 66.6 % switch rate).

In 33.3% of trials, the stop-signal (i.e., a yellow frame; Fig. 1a) indicated to participants that the planned action had to be cancelled. Stop-trials were counted as successful when participants did not make any responses within 800 ms time-window following the stop-signal onset. In Exp. 1, the stop-signal appeared 100 ms after the stimulus onset. In Exp. 2, the interval between the stimulus and stop-signal onset (i.e., the stop-signal delay or SSD) was adjusted using an adaptive staircase (tracking) method based on participants' trial-to-trial stopping success. Specifically, individuals' SSDs varied between 0 ms to 800 ms counting from the onset of the stimulus, starting from 100 ms of SSD at the beginning of session. Correct/incorrect stop trials increased/decreased SSDs by the step size that was randomly selected from 11.8 ms , 23.5 ms, or 35.3 ms for each trial. Go trials lasted until either the response was executed; stop trials lasted either until the 800 ms response window expired, or until (incorrect) response was emitted.

There were 2 practice blocks and 200 and 250 experimental blocks for Exp. 1 and 2 respectively. Each block lasted 16 seconds, within which participants were instructed to complete as many trials as possible; trials that were initiated began within the 16 second block duration were allowed to complete. The average number of go-trials and stop-trials were 1378 (SD = 91) and 685 (SD = 33) for Exp. 1, and 1576 (SD = 162) and 773 (SD = 75) for Exp. 2. Throughout the experimental session, participants were reminded to respond as accurately and fast as possible and to not “wait for” the stop-signal. In Exp. 2, participants were instructed that the tracking procedure would make it easier to stop on some trials and more challenging others. Participants were given a performance-based incentive for trials with RTs on go-trials faster than the 75th percentile of correct responses in the preceding blocks when 1) the overall accuracy in go-trials was above 90 percent and 2) there were more than 5 completed trials in a given block. While performing the task, participants were asked to rest the index finger in the center of the four response keys at the start of each trial (i.e., no lateralization of response sides). At the end of each trial, feedback (a green or red fixation cross) was presented based on the accuracy of responses in go-trials or on correct stopping in stop-trial. At the end of each block, the number of completed tests, the number of correct responses in go/stop-trials, and the amount of earned incentives, which reflects the speed of responses in go-trials, were presented as a feedback. All stimuli were created in Matlab (Mathworks) using the Psychophysics Toolbox (Brainard, 1997) and were presented on a 17-inch CRT monitor (refresh rate: 60 *Hz*) at a viewing distance of 100 cm.

***Stop-signal Reaction Time***

In Exp. 2, we computed individuals’ stop-signal reaction time (SSRT), which estimates the latent process of stopping, with the integration method 17. First, for each quantile bin of SSDs (Fig. 3b), the mean SSDs and the proportion of successful stop trials (p(respond|signal)) were calculated. Then, the matching go RTs were defined in each SSD bin by taking the nth RT in the rank ordered go-trial RTs (including all go-trials), where n is defined by multiplying

the number of RTs in the distribution by the probability of responding, p(respond|signal) or unsuccessful stopping, for each SSD bin. Within each SSD bin, SSRT was calculated by subtracting the corresponding SSD from the matching go RT, then scores from 6 SSD bins were averaged within individuals to obtain a single metric of SSRT for each individual. For all participants, failed-stop RTs were faster than correct go RTs.

EEG recordings and preprocessing

Scalp EEG activities were recorded from 20 tin electrodes on an elastic cap (Electro-Caps) using the International 10/20 system. The 10/20 sites F3, Fz, F4, T3, C3, CZ, C4, T4, P3, PZ, P4, T5, T6, O1, and O2 were used along with five nonstandard sites: OL halfway between T5 and O1; OR halfway between T6 and O2; PO3 halfway between P3 and OL; PO4 halfway between P4 and OR; and POz halfway between PO3 and PO4. Electrodes placed ~1cm to the left and right of the external canthi of each eye recorded horizontal electrooculogram (EOG) to measure horizontal saccades. To detect blinks, vertical EOG was recorded from an electrode placed beneath the left eye and reference to the left mastoid. The left-mastoid was used as reference for all recording sites, and data were re-referenced off-line to the average of all scalp electrodes. The scalp EEG and EOG were amplified with an SA Instrumentation amplifier with a bandpass of 0.01–80 *Hz*, and signals were digitized at 250 *Hz* in LabView 6.1 running on a PC. EEG data was first segmented by 18.5 second intervals to include all trials within a block. After time-frequency decomposition was performed (see Time-Frequency Analysis section), these epochs were further segmented into trial-to-trial epochs (the time interval of -600 to 800 ms relative to the onset of the stimulus for both experiments). These trial-to-trial epochs including blinks (>80 μv, window size = 200 ms, window step = 50 ms), large eye movements (>1°, window size = 200 ms, window step = 10 ms), blocking of signals (range = -0.01 μv to 0.01 μv, window size = 200 ms) were excluded from subsequent analyses.

### *Time-Frequency Analysis*

Temporal-spectral profiles of single-trial EEG data were obtained via complex wavelet analysis (Cohen, 2014) by applying time-frequency analysis to preprocessed EEG data epoched for each block (>18 seconds to exclude the edge artifacts). The power spectrum was convolved with a series of complex Morlet wavelets $e^{2\pi f t} e^{-t2/(2*\sigma^2)}$), where *t* is time, *f* is frequency increased from 1 to 35 *Hz* in 35 logarithmically spaced steps, and σ defines the width of each frequency band, set according to $n/2\pi f,$ where *n* increased from 3 to 10**.** The logarithmic scaling was used to keep the width across frequency band approximately equal, and the incremental number of wavelet cycles was used to balance temporal and frequency precision as a function of frequency of the wavelet. After convolution was performed in the frequency-domain, we took an inverse of the Fourier transform, resulting in complex signals in the time-domain. A frequency band-specific estimate at each time point was defined as the squared magnitude of the convolved signal $Z(real[z(t)]^2 + \ imag[z(t)]^2)$ for instantaneous power.

### *Representational Similarity Analysis*

The decoding analysis in the current study follows closely our previously established methods (Kikumoto and Mayr, 2020). We used a two-step procedure to obtain information about the strength of each action feature and conjunction on the level of individual trials and time samples within trials. First, we performed a linear decoding analysis to discriminate between all 12 action constellations. Specifically, we performed a penalized linear discriminant analysis using the caret package in R (Kuhn, 2008). At every time sample point, the instantaneous power of rhythmic EEG activity was averaged within the predefined ranges of frequency values (1-3 *Hz* for the delta-band, 4-7 *Hz* for the theta-band, 8-12 *Hz* for the alpha-band, 13-30 *Hz* for the beta-band, 31-35 *Hz* for the gamma-band), generating 100 features (5 frequency-bands X 20 electrodes) to train decoders. Within individuals, these data points were

z-transformed across electrodes at every sample to remove the effects that uniformly influenced all electrodes. We used a k-fold repeated cross-validation procedure to evaluate the decoding results (Mosteller and Tukey, 1968), by randomly partitioning single-trial EEG data into four independent folds. All trials except incorrect go-trials were used as the training sets in both experiments. The number of observations of each action constellation was kept equal within and across folds by dropping trials randomly. Three folds served as a training set and the remaining fold was used as a test set; this step was repeated until each fold served as a test set. Each cross-validation cycle was repeated eight times, in which each step generated a new set of randomized folds. Resulting classification probabilities (i.e., evidence estimated for each case of S-R mapping) were averaged across all cross-validated results with the best tuned penalty parameter. This decoding step yielded a vector of “confusion profiles” of classification probabilities for both the correct and all possible incorrect classifications and for each time point and trial (Fig. 1c).

As a second step, we then applied time-resolved RSAs to each profile of classification probabilities in order to determine their underlying similarity structure. Specifically, we regressed the confusion vector onto model vectors as predictors, which were derived from a set of representational similarity model matrices (Fig. 1c). Each model matrix uniquely represents a potential, underlying representation (e.g., rules, stimuli, responses and conjunctions). For example, the rule model predicts neural responses to be similar (i.e., more confusable) among instances of the same rule, but dissimilar across different rules. To estimate the unique variance explained by competing models, we regressed all model vectors simultaneously, which generated coefficients for each of the four model vectors. These coefficients (i.e., their corresponding t-values) allowed us to relate the dynamics of action representations to trial-to-trial variability in behavior in go- and stop-trials (see Multilevel Modeling section for details). In all RSAs, we logit-transformed classification probabilities and further included subject-specific “conjunction RT and stopping accuracy” models (i.e., two

vectors that each contained z-scored average RTs and stopping accuracy) as nuisance predictors to reduce potential biases in decoding due to idiosyncratic differences in performance among action constellations.

In both experiments, decoders were trained with the stimulus-aligned EEG signal. In Exp. 2, we further computed RSA scores that were re-epoched in reference to the onset of the stop-signal for every stop-trials with the variable SSDs (the right column of Fig. 4 and 5). Matching go-trial results were calculated by referencing the most updated SSDs for given trials that could have been used if the stop-signal appeared in those trials. We excluded resulting t-values that exceeded 5 SDs from means for each sample point, which excluded less than 1% of the entire samples in both experiments. Resulting t-values were averaged within 20 ms non-overlapping time samples. For decoding analyses and subsequent RSA with EEG, error-trials in go-trials were excluded.

***Estimating Timing of Stop-induced Suppression***

In Exp. 1, we used nonparametric permutation tests with a single- threshold method to identify the earliest time sample at which statistically significant differences between go-trials and stop-trials emerged. Specifically, for each action feature, we computed permutation distributions of the maximum statistic for every sample point from the stop-signal onset (fixed at 100 ms after the stimulus onset) to the end of 800 ms of the hold period. First, we obtained RSA results by decoding data with randomly shuffled condition labels (i.e., of action constellations). We then performed a series of t-tests, testing the differences in RSA scores between go- and stop-trials, for every sample against the null level (i.e., 0 for t-values). Out of the series of t-test results, we retained the maximum t-value. We repeated this process 10000 times by randomly drawing samples from all possible permutations of labels, thereby generating the permutation distributions of the maximum statistics. This approach allowed us to identify statistically significant, individual time points by comparing scores from the correct

labels to the critical threshold, which was defined as the 95th (i.e., alpha =.05) of the largest member of maximum statistics in the permutation distribution of the corresponding variable.

***Multilevel Modeling***

In Exp. 2, to analyze predictors of trial-by-trial variability in stopping success, we used multilevel logistic regression models. Specifically, we estimated for stop trials a model predicting stopping success on a given trial using the RSA-derived t-values for basic action features (i.e., rule, stimulus, and response) and the conjunction as predictors. In addition, we also included each trial's log-transformed SSD as a covariate to account for the possibility that SSDs affect both action representations and stopping success as a third-variable. For statistical tests, we used EEG data averaged over two a-priori selected, symmetric time intervals: the pre-stop-signal period (-200 to 0 ms) and the post-stop-signal period (0 ms to 200 ms), in reference to the onset of the stop-signal in each trial. Both time intervals clearly precede the average SSRT across individuals (M = 272 ms). We also performed additional control analyses, where we excluded trials with early responses (i.e., responses occurred after the stimulus onset and before stop-signal in unsuccessful-stop trials) and where we included decoded representations from both pre-/post-stop-signal simultaneously (Table 3). In addition, to visualize changes in predictability of stopping success, we separately performed a series of logistic regression analyses by fitting models at each sample point in reference to the onset of the stimulus and the stop-signal (Fig. 5). To assess how action representations contributed to action selection in go-trials, we also report for both experiments results from multilevel models to assess which action representations predict trial-to-trial RTs on go trials (Table 2). Here, RTs were log-transformed and trials with response errors were excluded.

## Result

### *Experiment 1*

#### *Behavior*

Behavioral performance in go-/stop-trials are summarized in Table 1. Average RTs and errors in go-trials were similar to the previous results using a paradigm with no stopping requirements (Kikumoto and Mayr, 2020). The probability of stopping failures—incorrectly executing responses in the presence of the stop-signal (i.e., *p*(respond|signal))—was low because of the early presentation of stop-signal at the fixed timing (100 ms after the stimulus onset), which allowed us to estimate the time-course of suppression of action representations from a fixed starting point.

#### *Action Representations in Go-trials and Stop-trials*

Fig. 2 shows the time-course of RSA scores estimated on the level of single trials for each of the basic features (i.e., rules, stimuli, and responses) and the conjunction, and for both go- and stop-trials. For go-trials, the flow of activated representations was highly consistent with our previous results: rule information appeared in the pre-stimulus period, and shortly after stimulus information peaked, response information emerged (Hubbard et al., 2019; Kikumoto and Mayr, 2020). Importantly, conjunctive information was present throughout the entire response-selection period. We also replicated our findings that trial-to-trial variability in the conjunctive representations robustly predicted go-trial RTs (Table 2), over and above other representations of constituent features, indicating the critical role of integrated representations for the execution of goal-directed actions.

Our main goal in Exp. 1 was to test the prediction that conjunctive representations are suppressed on stop-trials relative to go-trials. Indeed, we found stopping of actions markedly reduced the strength of conjunctive representations right after the onset of the stop-signal (Fig. 2). Not surprisingly, the response representation was also suppressed, whereas we found no effect on the rule representation and only a relatively late, and small effect for the stimulus

representation. Importantly, suppression of the conjunction occurred at the same time, or even slightly before suppression of the response representation. This suggests that the reduction of conjunctive information is not just an aftereffect of response suppression. Rather, it supports the notion that the conjunctive representation is a direct target of the stopping activity.

In addition, the inserts in Fig. 2 show scores when only the simple features are used in the RSA model. Here, the suppression effect on the basic features, specifically for the rule representation was substantially increased, highlighting the importance of including the conjunction model (see also, Kikumoto and Mayr, 2020).

***Experiment 2***

The results of Exp. 1 are consistent with the hypothesis that conjunctive representations are a primary target of the stopping process. Yet, in principle, such a suppression effect could simply be an epiphenomenon of stopping-related events, such as the surprise from an unexpected stop-signal. Therefore, it would be important to establish a functional relationship between the suppression effect on the conjunctive representations on the one hand and stopping behavior on the other. In Exp. 2, we therefore attempted to link trial-to-trial variability in action representations to trial-to-trial variability in stopping success. For this purpose, we used the standard, stop-signal paradigm with an adaptive staircase algorithm. By adjusting the stop-signal delay within-subjects, we achieved approximately equal numbers of successful and failed stopping in the presence of the stop-signal (see sections *Stimuli, Tasks and Procedure* and *Stop-signal Reaction Time* for details).

***Behavior***

Behavioral performance is summarized in Table 1. Most participants (33 out of 36 participants) exhibited *p*(stop|signal) in the range of .40 - .65 (individuals with the stopping accuracy higher than 75% were excluded from further analyses). The average RTs in go-trials were longer than the RT in failed stop-trials for all participants, indicating the validity of the race model to estimate individuals' stop-signal reaction time (Fig. 3a). The probability of stopping

errors (i.e., inhibition functions) covaried with the increase of SSDs, indicating the efficacy of the SSD staircase algorithm (Fig. 3b).

***Action Representations in Go-trials, Failed Stop-trials, and Successful Stop-trials***

Fig. 4 shows RSA scores for the individual features and the conjunction that are aligned to the onset of the stimulus (left column) and the stop-signal (right column), separately for go-trials, failed stop-trials, and successful stop-trials. Again, we found that rule and stimulus representations showed little effect of stopping (but excluding the conjunction model also yielded similar results to Exp.1). In contrast, we found that conjunctive representations, and late response representations, were selectively suppressed in successful stop-trials relative to go-trials and failed stop-trials. For the conjunctive representations, the divergence seemed to occur even before the onset of the stop-signal (see individuals' average SSDs in Fig. 4 left column), suggesting stopping was particularly impaired when the conjunctive representations developed strongly in the early response selection phase. Indeed, when RSA scores are replotted relative to trial-to-trial SSDs, differences in successful and failed stop-trials emerged clearly before the average SSRT (M=272 ms) and even slightly before the stop-signal. No other action features showed the suppression effects in the pre-stop-signal period ($t < .13$), and post-stop-signal effects on the response representation were apparent only when the conjunction model was excluded, $b=-.010$, $SE=.004$, $t(33)=2.34$. This result is consistent with the prediction that the strength of the conjunction at the time the stop process is initiated, determines stopping success. We also note that as in Exp. 1 and in our previous work (Kikumoto and Mayr, 2020), the strength of conjunctive representations was also the most robust predictor of trial-to-trial RTs for go-trials (Table 2).

***Predicting Successful Stopping by Decoded Representations***

Results so far indicate that the state of conjunctive representations prior to the onset of the stop-signal determines the success of stopping. This predictive, pre-stop-signal effect needs to be further confirmed by establishing its robustness against the influence of other

action representations, potential third-variables that covary with the development of conjunctions (e.g., trial-to-trial SSDs), and autocorrelation of signals in case of slow oscillations. To this end, we performed multilevel logistic regressions to predict single-trial stopping failures using decoded action features and SSDs as simultaneous predictors. Critically, we found that the state of conjunctive representations, over and above the state of the other features, strongly predicts single-trial stopping failures prior to the onset of the stop-signal (Fig. 5). These results were robust when we accounted for the trial-to-trial SSDs or the premature responses that occurred before the stop-signal onset (see Fig. 5 and Table 3). In addition, the conjunctive representations in the pre-stop-signal phase (-200 to 0 ms) and the post-stop-signal phase (0 to 200 ms) uniquely predicted stopping success when both effects were included in the same model (Table 2).

## Discussion

Even simple, goal-directed actions, rely on various aspects of the task environment. Event-file theory proposes that different action features need to be integrated within conjunctive representations to guide actions (Hommel et al., 2001; Rigotti et al., 2013). In our previous work (Kikumoto and Mayr, 2020), we had reported EEG-decoding evidence for representations that behave like event files. In the current work, we tested the hypothesis that because such representations are critical for action control, they should also be intricately involved in the stopping of a planned, or initiated action. Specifically, we predicted that conjunctive representations are the main target of the stopping process and that the strength of conjunctive representations at the time the stopping process is initiated, inversely predicts stopping success. Our results fully confirmed these predictions: Conjunctive representations were selectively suppressed in response to the stop-signal (Fig. 2), and stopping became more challenging on trials with strong conjunctive representations, right before the arrival of the stop-signal (Fig. 4 and Fig. 5).

Studies of action inhibition typically compare the “go” and successful/failed “stop” processes in an aggregated manner. This leaves the question, which action-related representations are influenced by stopping. Theoretically, stopping of actions may require suppression of all task-relevant representations simultaneously. Alternatively, only those representations directly involved with motor control might be targeted. Instead, our results suggest that the conjunctive representation is the primary target of suppression, followed, not surprisingly, by the response representation (Fig. 2 and Fig. 4 and Fig. 5). It is an open question whether conjunctive and response representations are separately targeted, or whether the deactivation of response representations is an aftereffect of the suppression of conjunctive representations. Representations of the rule or the stimulus remained intact, or showed very minor suppression, only after the completion of the stopping process (Fig. 4). In particular, the representation of the rule was unaffected by the stopping process. A potential functional benefit of selective suppression in real-world situations is that by retaining the abstract rule information, actions can be easily re-implemented, once the reason for stopping has been removed.

Conjunctive representations that integrate disparate features, including abstract rule information, are by definition situated on a more central level than representations that directly control motor output. The fact that conjunctive representations were targeted by the stopping process, is consistent with recent results indicating that inhibition of actions and inhibition of thoughts or memories are handled by a shared process (Anderson, 2004; Guo et al., 2018). For example, using the Think/No-think paradigm, studies found that the same right lateral prefrontal area that is typically involved in stopping of motor responses, was also critical in suppressing involuntary intrusions of retrieved thoughts. An interesting question for future research is whether the fact that conjunctions seem to be a key target of inhibition has implications for the suppression-induced, long-term memory effects (Anderson and Green, 2001). In principle, conjunctive representations provide contextual specificity to more abstract

feature codes for a given task. Thus, as the main target of inhibitory control they may help constrain otherwise, unmitigated spread of inhibition beyond the current context.

Our results showed that the pre-stop-signal state of the conjunctive representation uniquely predicts the success of subsequent stopping, over and above the potential effects of a reactive inhibition process that is initiated after the stop-signal (Table 3). This pattern strongly suggests that the strength of conjunctive representations is a key driver of the efficiency and success of the action—and therefore also of the ability to stop that action. However, our results by themselves do not identify the mechanisms that modulate the state of the conjunctive representation prior to the stop-signal. One possibility is that conjunction strength depends on endogenous fluctuations of attention towards the go action across trials. In fact, a strong emphasis on initiating action may induce strong conjunctions and thereby cause the failures to trigger the stop process altogether (Matzke et al., 2017b; Matzke et al., 2017a). The fact that conjunctive representations in failed-stop trials, prior to the arrival of stop-signal, were even stronger than on go-trials, is consistent with such an attentional fluctuation account. As another possibility, there is evidence that variations in strategic, proactive inhibition affect the state of conjunctions (Aron, 2011). On trials in which subjects anticipate stopping, proactive inhibition may keep conjunctive representations from fully developing. Such proactive control processes could be directly tested, by cuing the stop probability on a trial-by-trial basis (Zandbelt et al., 2013; Vink et al., 2015). In any case, our results clearly indicate that the pre-stop-signal state of action representations must be taken into account to fully understand subsequent reactive inhibition and stopping.

One potential caveat is that for the task space used in Exp. 1 and 2, the RSA analyses allow us to say with certainty that at least two different task-relevant features were integrated within conjunctions, but do not allow us disambiguate between conjunctions that include binary combinations of rules, stimuli, or responses, or the combination between all three aspects. However, in our previous work we had used--aside from the current task space (Exp.1 in

Kikumoto and Mayr, 2020)--also an expanded task space with a more complex set of rules, in which we were able to disambiguate between different types of conjunctions (Exp.2 in Kikumoto and Mayr, 2020). In terms of functional characteristics, the conjunctions derived from the limited task space and the most complex conjunctions derived from the expanded task space behaved in a highly similar manner. Therefore, we are confident that in the design we used here, the conjunctions reflect an integration between both the rule, and stimulus/response features.

Our EEG-based decoding results provide no precise information about the neural-anatomical location of conjunctive representations (for further explorations of the neural basis of these representations (for exploratory analyses of the physiological basis of these representatinos, see supplementary information to Kikumoto and Mayr, 2020). However, recently, there has been increasing evidence from research with non-human primates about the high prevalence of neurons in parietal/frontal areas that show very similar properties as the conjunctive representations we report on (Barak et al., 2013; Rigotti et al., 2013; Fusi et al., 2016; Parthasarathy et al., 2017). Specifically, these so-called mixed-selectivity neurons integrate rule, stimulus, and response information in a nonlinear manner and, as the EEG-decoded conjunctive representations, are highly predictive of observed actions, over and above their constituent features. It would be important to establish to what degree the representations examined here, are a reflection of mixed-selectivity, neuronal activity. One way to test this hypothesis is to look for equivalent functional relationships in both human and animal models, such as between the strength of conjunctive activity in animal models and attempts to stop the action plan supported by this activity.

In conclusion, the results we report here build on our previous work suggesting that conjunctive, event-file type representations can be tracked with high temporal resolution through EEG-decoding techniques and can be shown to be highly relevant for trial-to-trial variation in behavior.  Specifically, our results are consistent with the hypothesis that such

conjunctive representations are a prime target of a putative action inhibition process, exactly because they are a key driver of successful action implementation.

## References


Anderson MC (2004) Neural Systems Underlying the Suppression of Unwanted Memories. Science 303:232-235.

Anderson MC, Green C (2001) Suppressing unwanted memories by executive control. Nature 410:366-369.

Aron AR (2011) From reactive to proactive and selective control: developing a richer model for stopping inappropriate responses. Biol Psychiatry 69:e55-68.

Aron AR, Robbins TW, Poldrack RA (2014) Inhibition and the right inferior frontal cortex: one decade on. Trends Cogn Sci 18:177-185.

Barak O, Rigotti M, Fusi S (2013) The sparseness of mixed selectivity neurons controls the generalization-discrimination trade-off. J Neurosci 33:3844-3856.

Brainard DH (1997) The psychophysics toolbox. Spat Vis 10:433-436.

Cohen MX (2014) Analyzing Neural Time Series Data: Theory and Practice: MIT Press.

Coxon JP, Stinear CM, Byblow WD (2006) Intracortical inhibition during volitional inhibition of prepared action. J Neurophysiol 95:3371-3383.

Donders FC (1969) On the speed of mental processes. Acta Psychol 30:412-431.

Duque J, Greenhouse I, Labruna L, Ivry RB (2017) Physiological Markers of Motor Inhibition during Human Behavior. Trends Neurosci 40:219-236.

Fusi S, Miller EK, Rigotti M (2016) Why neurons mix: high dimensionality for higher cognition. Current opinion in neurobiology 37:66-74.

Greenhouse I, Sias A, Labruna L, Ivry RB (2015) Nonspecific Inhibition of the Motor System during Response Preparation. J Neurosci 35:10675-10684.

Guo Y, Schmitz TW, Mur M, Ferreira CS, Anderson MC (2018) A supramodal role of the basal ganglia in memory and motor inhibition: Meta-analytic evidence. Neuropsychologia 108:117-134.

Hommel B (1998) Event files: Evidence for automatic integration of stimulus-response episodes. Vis Cogn 5:183-216.

Hommel B (2019) Theory of Event Coding (TEC) V2.0: Representing and controlling perception and action. Atten Percept Psychophys.

Hommel B, Müsseler J, Aschersleben G, Prinz W (2001) The Theory of Event Coding (TEC): a framework for perception and action planning. Behav Brain Sci 24:849-878; discussion 878-937.

Hubbard J, Kikumoto A, Mayr U (2019) EEG Decoding Reveals the Strength and Temporal Dynamics of Goal-Relevant Representations. Sci Rep 9:9051.

Kikumoto A, Mayr U (2020) Conjunctive representations that integrate stimuli, responses, and rules are critical for action selection. Proceedings of the National Academy of Sciences:201922166.

Kornblum S, Hasbroucq T, Osman A (1990) Dimensional overlap: cognitive basis for stimulus-response compatibility--a model and taxonomy. Psychol Rev 97:253.

Kriegeskorte N, Mur M, Bandettini P (2008) Representational similarity analysis - connecting the branches of systems neuroscience. Front Syst Neurosci 2:4.

Kuhn M (2008) Caret package. Journal of Statistical Software 28(5).

Labruna L, Lebon F, Duque J, Klein P-A, Cazares C, Ivry RB (2014) Generic inhibition of the selected movement and constrained inhibition of nonselected movements during response preparation. J Cogn Neurosci 26:269-278.

Logan GD, Cowan WB (1984) On the ability to inhibit thought and action: A theory of an act of control. Psychol Rev 91:295.

Matzke D, Love J, Heathcote A (2017a) A Bayesian approach for estimating the probability of trigger failures in the stop-signal paradigm. Behav Res Methods 49:267-281.

Matzke D, Hughes M, Badcock JC, Michie P, Heathcote A (2017b) Failures of cognitive control or attention? The case of stop-signal deficits in schizophrenia. Atten Percep Psychophys 79:1078-1086.
Mayr U, Bryck RL (2005) Sticky rules: integration between abstract rules and specific actions. J Exp Psychol Learn Mem Cogn 31:337-350.
Mosteller F, Tukey JW (1968) Handbook of Social Psychology. 2:80-203.
Parthasarathy A, Herikstad R, Bong JH, Medina FS, Libedinsky C, Yen S-C (2017) Mixed selectivity morphs population codes in prefrontal cortex. Nat Neurosci 20:1770-1779.
Posner MI, Mitchell RF (1967) Chronometric analysis of classification. Psychol Rev 74:392.
Rigotti M, Barak O, Warden MR, Wang X-J, Daw ND, Miller EK, Fusi S (2013) The importance of mixed selectivity in complex cognitive tasks. Nature 497:585.
Sanders AF, Sanders A (2013) Elements of human performance: Reaction processes and attention in human skill: Psychology Press.
Schumacher EH, Hazeltine E (2016) Hierarchical task representation: Task files and response selection. Curr Dir Psychol Sci 25:449-454.
Sternberg S (1969) The discovery of processing stages: Extensions of Donders' method. Acta Psychol 30:276-315.
Swann N, Tandon N, Canolty R, Ellmore TM, McEvoy LK, Dreyer S, DiSano M, Aron AR (2009) Intracranial EEG reveals a time- and frequency-specific role for the right inferior frontal gyrus and primary motor cortex in stopping initiated responses. J Neurosci 29:12675-12685.
Verbruggen F, Aron AR, Band GP, Beste C, Bissett PG, Brockett AT, Brown JW, Chamberlain SR, Chambers CD, Colonius H (2019a) A consensus guide to capturing the ability to inhibit actions and impulsive behaviors in the stop-signal task. Elife 8:e46323.
Verbruggen F et al. (2019b) A consensus guide to capturing the ability to inhibit actions and impulsive behaviors in the stop-signal task. Elife 8.
Vink M, Kaldewaij R, Zandbelt BB, Pas P, du Plessis S (2015) The role of stop-signal probability and expectation in proactive inhibition. Eur J Neurosci 41:1086-1094.
Wessel JR (2019) β-bursts reveal the trial-to-trial dynamics of movement initiation and cancellation. J Neurosci.
Zandbelt BB, Bloemendaal M, Neggers SFW, Kahn RS, Vink M (2013) Expectations and violations: delineating the neural network of proactive inhibitory control. Hum Brain Mapp 34:2015-2024.

Figures

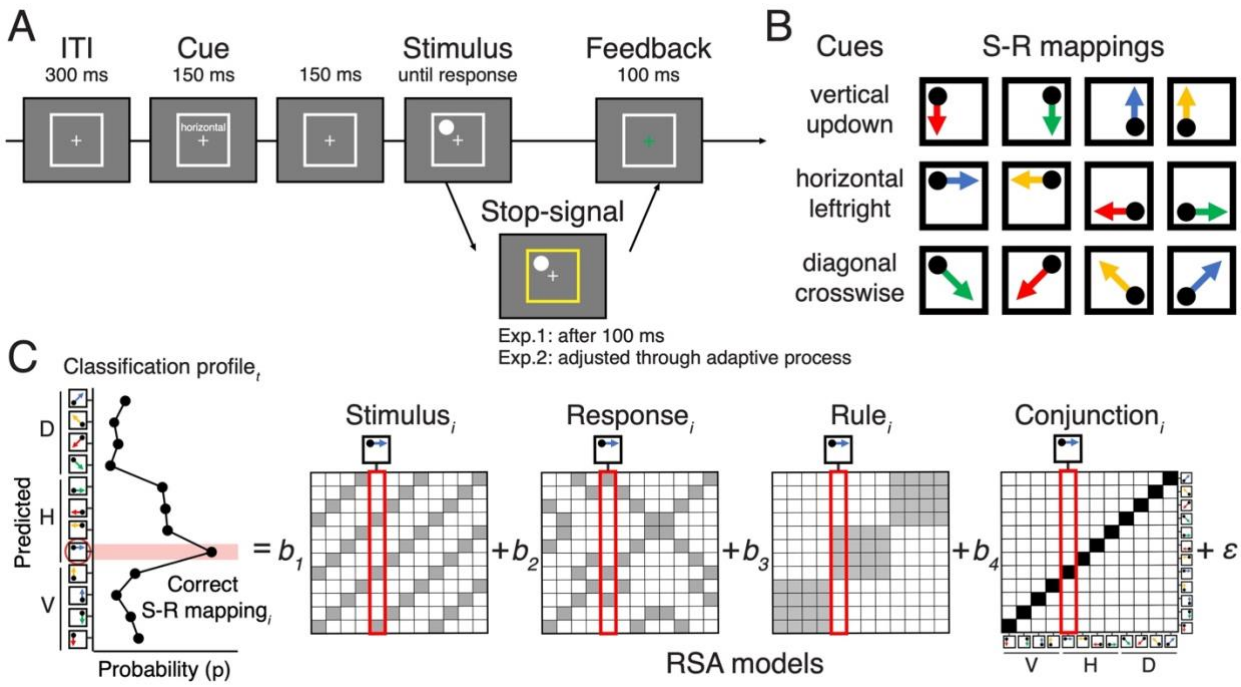


**Fig. 1. Task design and analytic procedure.**
(A) Sequence of trial events in the combined rule-selection/stop-signal task for both Exp. 1 and 2. (B) Spatial translation rules mapping specific stimuli to responses. Two different cue words were used for each rule to disambiguate between cue and rule-level representations. (C) Schematic steps of the representational similarity analysis. The raw EEG signal was decomposed into frequency-band specific activity via time-frequency analysis (see *EEG recordings and preprocessing* and *Time-Frequency Analysis*). For each sample time (*t*), a scalp-distributed pattern of EEG power was used to decode the specific rule/stimulus/response configuration of a given trial, producing a set of classification probabilities for each of the possible configurations. The profile of classification probabilities reflects the similarity structure of the underlying representations, where similar action constellations are more likely to be confused. The example of profile of classification probabilities shows the case where a unique conjunction and rule information is expressed (peak at the correct S-R mapping$_i$ and confusion to other instances with the same rule). For each trial and timepoint, the profile of classification probabilities is simultaneously regressed onto model vectors as predictors that reflect the different, possible representations. In each matrix of model vectors, the x-axis corresponds to the correct constellation for the decoder to pick, and the y-axis shows all possible constellation. The shading of squares indicates the predicted classification probabilities (darker shading means higher probabilities). The coefficients associated with each predictor (i.e., *t*-values) reflect the unique variance explained by each of the constituent features and their conjunction.

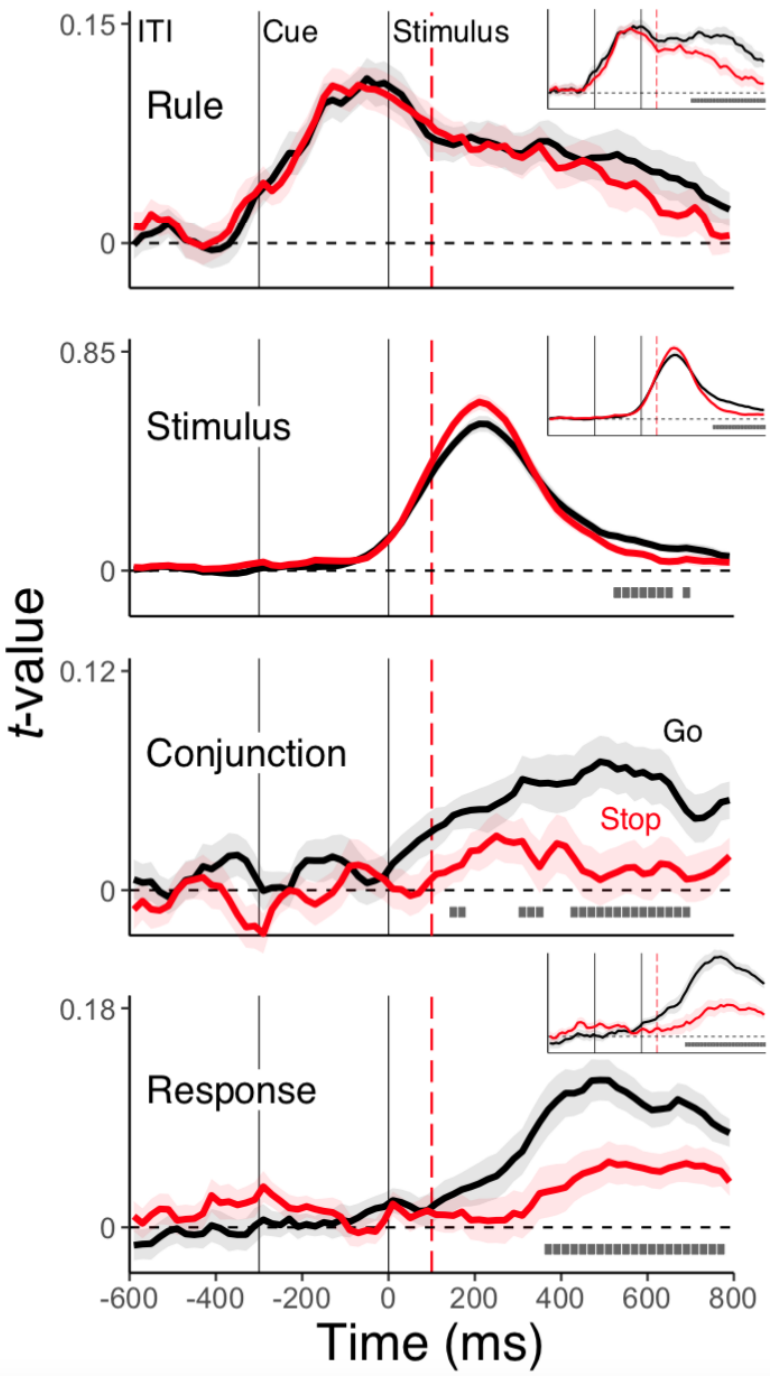


**Fig 2. Effects of stopping on the expression of each feature in the RSA analysis for Exp. 1.**

Average, single-trial *t*-values derived from the RSA (see Fig. 1C) for each of the basic features (rule, stimulus, and response) and the conjunction, separately for go-trials (black) and stop-trials (red). Shaded regions specify the 95% within-subject confidence intervals. The vertical, red dashed line marks the onset of the stop-signal at 100 ms after the stimulus onset. Squares below lines denote the time points with significant differences between go- and stop-trials, correcting for multiple comparison using a non-parametric permutation test. The inserts for the rule, stimulus, and response features show the same results when the RSA contains only these basic features, but excludes the conjunction as model predictor.

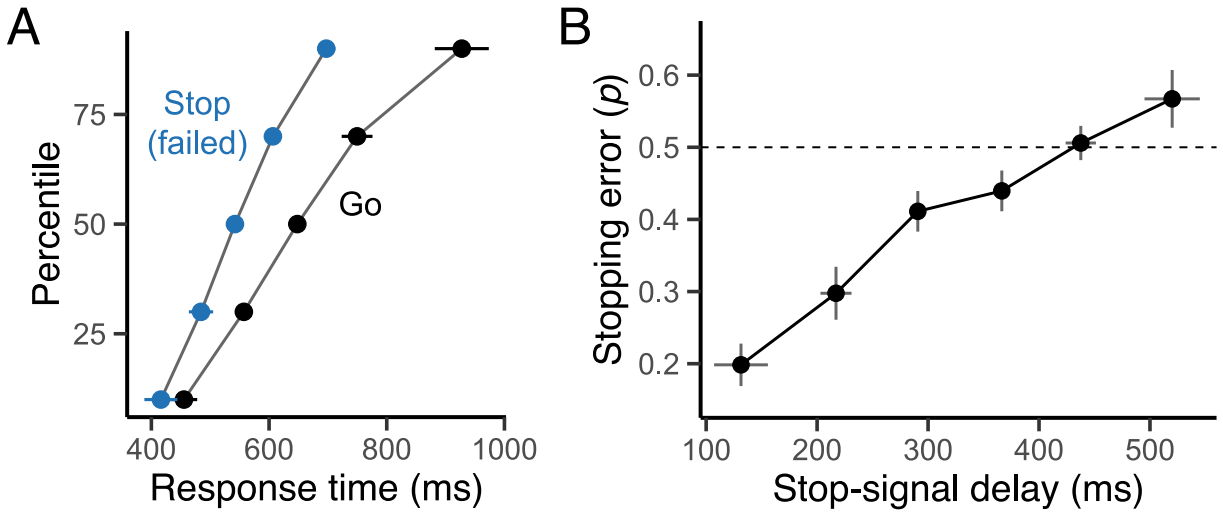


**Fig. 3**. **Behavioral results for Exp. 2.**
(A) Vincentized mean response times (RTs) for go-trials and failed stop-trials. (B) Average rates of stopping failures as a function of stop-signal delays. Error bars specify 95% within-subject confidence intervals.

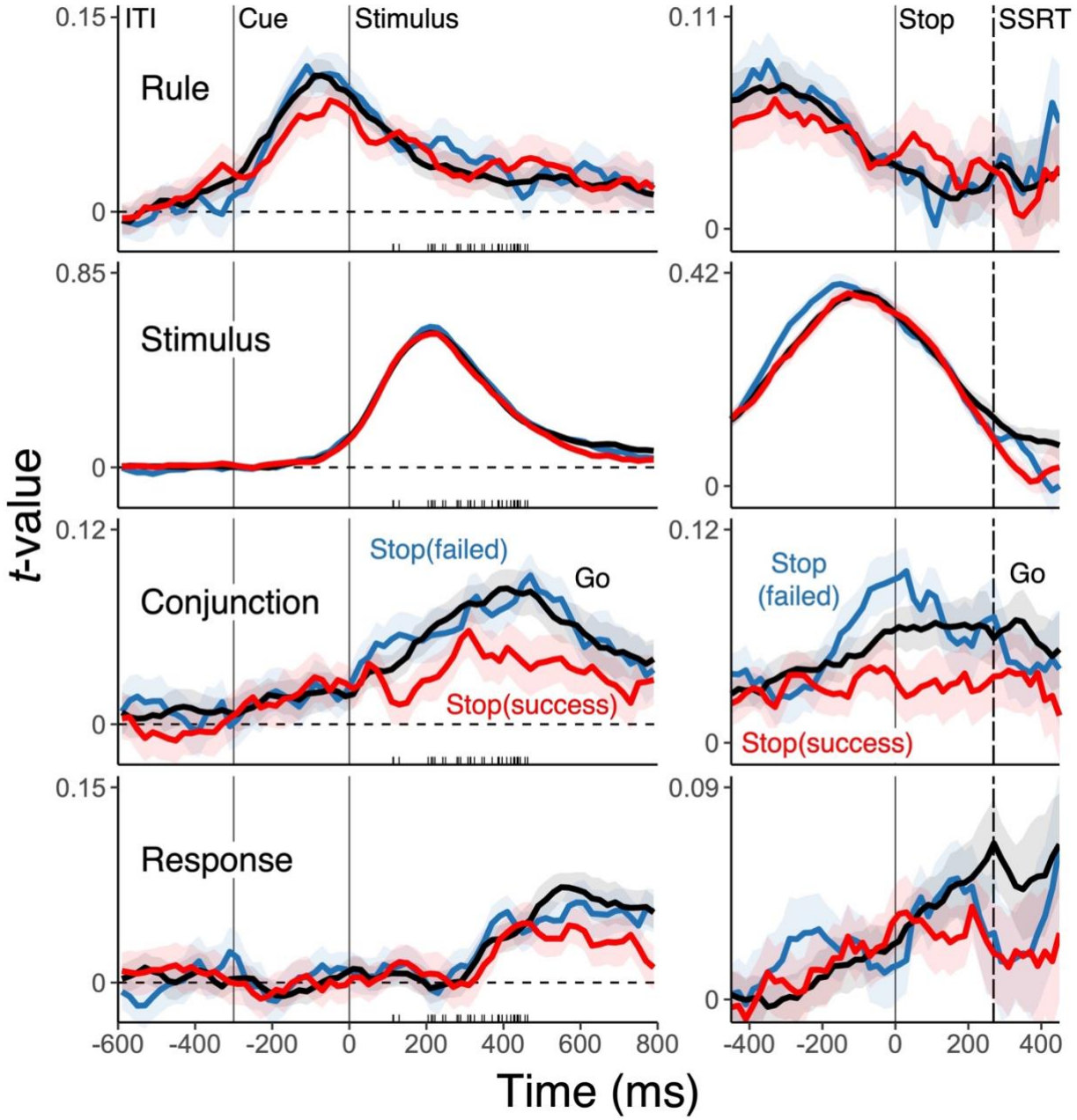


**Fig 4. Effects of stopping on the expression of each feature in the RSA analysis for Exp. 2.**

Average, single-trial *t*-values associated with each of the basic features (rule, stimulus, and response) and their conjunction derived from the RSA, separately for go-trials (black), successful stop-trials (red), and failed stop-trials (blue). The left panels show the results aligned with the stimulus onset, the right panels with the stop-signal onset. Shaded regions specify the 95% within-subject confidence intervals. Tick marks on the x-axis mark individuals' average stop-signal delays.

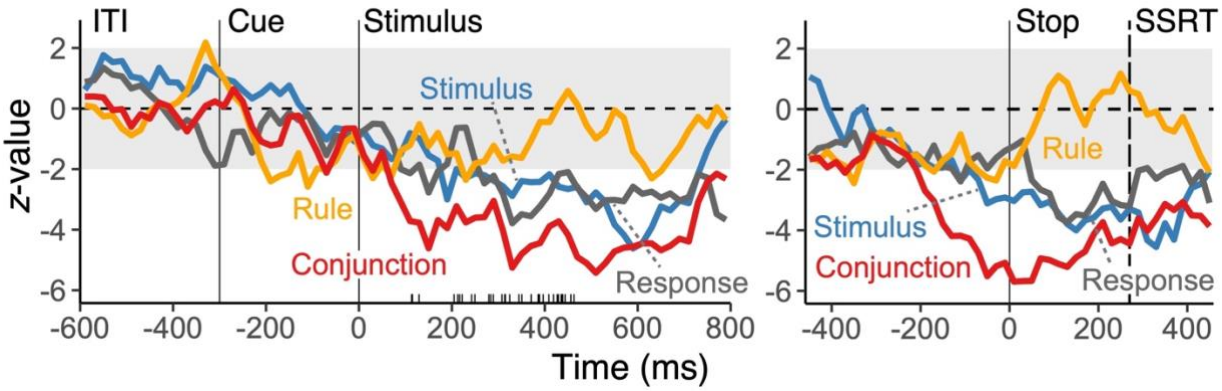


**Fig. 5. Predicting stopping success from strength of representations.**
Time-course of *z* values from multilevel, logistic regression models predicting the variability in trial-to-trial stopping failures in the stop-trials (the "impact" of representations on stopping success), using RSA scores of all features and trial-to-trial SSDs as simultaneous predictors. Negative *z*-value indicates more stopping failures as the strength of decoded representations increase. The left panel shows results aligned to the stimulus onset; in the right panel data are aligned to the stop-signal onset.

## Tables

*Table 1*. Behavioral performance in go- and stop-trials.

| | Exp.1 | Exp.2 |
|---|---|---|
| RT (Go, ms) | 503 (62.5) | 675 (139) |
| Error (Go, %) | 3.24 (1.77) | 3.55 (3.64) |
| *p*(respond\|signal) | 11.6 (8.85) | 44.1 (4.84) |
| Failed Stop RT (ms) | 405 (32.2) | 552 (91) |
| Error (Stop, %) | 2.36 (2.66) | 2.83 (2.28) |
| SSD (ms) | | 327 (103) |
| SSRT (ms) | | 272 (54.8) |

Note. Each column corresponds to 1) the average RT in go-trials (Go RT) excluding response errors, 2) the rate of response errors in go-trials, 3) the rate of unsuccessful stopping (p(respond | signal)) in stop-trials, 4) the average RT of incorrectly executed responses in stop-trials, and 5) the rate of response errors in incorrectly executed responses in stop-trials.

*Table 2*. Predicting trial-by-trial RTs in go-trials using the average strength of representations decoded through the RSA analyses during 0-300 ms post-stimulus intervals for each trial.

| | Exp.1 | | Exp.2 | |
|---|---|---|---|---|
| Variable | *b* (*se*) | *t*-value | *b* (*se*) | *t*-value |
| Rule | -.025 (.012) | -2.15 | -.013(.006) | -2.08 |
| Stimulus | -.015 (.009) | -1.59 | -.038 (.009) | -4.12 |
| Response | -.016 (.009) | -1.85 | -.032 (.009) | -3.55 |
| Conjunction | -.042 (.012) | -3.61 | -.057 (.010) | -5.69 |

Note. Coefficients for all decoded variables were included as predictors simultaneously. Negative coefficient imply that stronger representations predict faster RTs.

*Table 3*. Predicting trial-by-trial stopping accuracy using the strength of decoded representations in Exp. 2.

| | | Pre-Stop-Signal | | Post-Stop-Signal | |
|---|---|---|---|---|---|
| Model | Variable | *b* (*se*) | *t*-value | *b* (*se*) | *t*-value |
| SSD control | Rule | -.077 (.033) | -2.38 | -.012 (.024) | -0.49 |
| | Stimulus | -.083 (.033) | -2.50 | -.060 (.020) | -3.11 |
| | Response | -.081 (.034) | -2.41 | -.062 (.020) | -3.11 |
| | Conjunction | -.193 (.041) | -4.70 | -.151 (.029) | -5.23 |
| Exclude early responses | Rule | -.050 (.020) | -2.43 | -.019 (.025) | -0.76 |
| | Stimulus | -.036 (.017) | -2.19 | -.057 (.020) | -2.89 |
| | Response | -.039 (.021) | -1.86 | -.054 (.020) | -2.70 |
| | Conjunction | -.136 (.029) | -4.78 | -.155 (.029) | -5.31 |
| Pre/post stop-signal simultaneous | Rule | -.084 (.036) | -2.31 | .016 (.044) | 0.36 |
| | Stimulus | -.035 (.036) | -0.93 | -.104 (.041) | -2.53 |
| | Response | -.038 (.040) | -0.97 | -.093 (.041) | -2.30 |
| | Conjunction | -.126 (.049) | -2.60 | -.194 (.049) | -3.93 |

Note. The “SSD control” model included trial-to-trial stop-signal delays as the fixed and random effect as covariate. In the “exclude early responses” model all premature responses that occurred prior to the onset of the stop-signal were removed. Whereas these models were fitted separately for pre-stop-signal and post-stop-signal predictors, in the “pre/post stop signal simultaneous” model, both pre-stop-signal and post-stop-signal predictors were included simultaneously. Pre-stop-signal interval: -200–0 ms; post-stop-signal interval: 0-200 ms.