CONJUNCTIVE REPRESENTATIONS IN DYNAMIC ACTION CONTROL

by

ATSUSHI KIKUMOTO

A DISSERTATION

Presented to the the Department of Psychology
and the Graduate School of the University of Oregon
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

March 2020

DISSERTATION APPROVAL PAGE

Student: Atsushi Kikumoto

Title: Conjunctive Representations in Dynamic Action Control

This dissertation has been accepted and approved in partial fulfillment of the requirements for the Doctor of Philosophy degree in the Department of Psychology by:

Ulrich Mayr	Chairperson
Brice Kuhl	Core Member
Sarah Dubrow	Core Member
Ian Greenhouse	Institutional Representative

and

Kate Mondloch	Interim Vice Provost and Dean of the Graduate School

Original approval signatures are on file with the University of Oregon Graduate School.

Degree awarded March 2020

© 2020 Atsushi Kikumoto

DISSERTATION ABSTRACT

Atsushi Kikumoto

Doctor of Philosophy

Department of Psychology

January 2020

Title: Conjunctive Representations in Dynamic Action Control

In the present work, I examine the functional role of highly integrated, conjunctive representations of basic task features during dynamic action control. People can use abstract action rules to flexibly configure and select actions for specific situations. Yet, how exactly rules shape actions towards specific sensory and/or motor requirements remains unclear. One theoretical possibility is that rules become integrated with sensory/response features in a nonlinear, conjunctive manner during action selection (i.e., event files). Such conjunctive representations in turn are a precursor of successful action. To test this hypothesis, it is necessary to dynamically track neural representation of multiple action features that become active concurrently during action selection. We applied multivariate decoding analysis to the time-resolved EEG signal at the level of single-trials, while participants selected actions based on varying action rules. In Chapter II, we provide initial evidence that conjunctive representations can be tracked during action selection. Specifically, we show that these representations emerged throughout the entire response selection period and that they were robust and unique predictors of the variability in trial-to-trial performance. Moreover, they were related to a theoretically important, behavioral indicator of event files—the partial-overlap priming effects. In Chapter III, we tested how conjunctive representations contribute to stopping of planned actions. Because

the formation of conjunctive representations is theorized as a necessary stage of successful actions, we hypothesized conjunctions should be the primary target of stopping-related activity. Indeed, using the stop-signal paradigm, we found (a) that conjunctions were selectively suppressed on stop trials, and (b) that stopping became particularly difficult when the conjunctive representations of to-be-stopped actions prior to the stop-signal. In Chapter IV, we discuss implications of the findings and propose ideas for future directions. Specifically, I propose further experiments to characterize key processes or properties associated with conjunctive representations, such how they are actively maintained in working memory, how they are functionally related to the dimensionality of neural responses, or how effects of actions are integrated in such representations. The work presented in this dissertation confirms that conjunctive representations are functionally independent of the constituent features and play a critical role in action control. It also provides broad insights to how we can study cognitive control functions in humans by directly decoding goal-relevant information. Chapter II and III are co-authored materials with Ulrich Mayr.

CURRICULUM VITAE

NAME OF AUTHOR: Atsushi Kikumoto

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene, OR
Lane Communicy College, Eugene, OR

DEGREES AWARDED:

Doctor of Phiolsopy, Psychology, 2020, University of Oregon
Master of Science, Psycholohy, 2014, University of Oregon
Bachelor of Science, Psychology, 2011, University of Oregon

AREAS OF SPECIAL INTEREST:

Cognitive Neuroscience
Cognitive Control

PROFESSIONAL EXPERIENCE:

Teaching Assistant, University of Oregon, 2013-2019

Junior Fellow, Max Planck Institute of Human Development, 2018
Cognitive and Neuronal Dynamic of Memory Across Lifespan

GRANTS, AWARDS, AND HONORS:

Pre-doctroral stipend for a junior visiting scholar, Max Planck Society, 2018

Gregores Resarch Award, University of Oregon, 2017

Graduate Student Award, Cognitive Neuroscience Society, 2015

International Deans Excellence Award Scholarship, University of Oregon, 2009-2012

Best Poster Presentation Award, The Northwest Cognitive & Memory (NOWCAM), 2011

Shining Star Scholarship, Lane Community College, 2008

PUBLICATIONS:

Kikumoto, A., & Mayr, U. (2019). Balancing model-based and memory-free action selection under competitive pressure. *eLife*, *8*.

Moss, M., Kikumoto, A. & Mayr, U. (in press). Does conflict regulation rely on working memory?. *Journal of Experimental Psychology: Learning, Memory and Cognition.*

Sereno, M.E., Robles, K.E., Kikumoto, A. & Bies, A.J. (in press). The Effects of 3-Dimensional Context on Shape Perception. *Psychological Science.*

Hubbard, J.*, Kikumoto, A*., & Mayr, U. (2019). EEG Decoding Reveals the Strength and Temporal Dynamics of Goal-Relevant Representations. *Scientific Reports*, *9*(1), 9051. * shared first-authorship.

Kikumoto, A., & Mayr, U. (2018). Decoding hierarchical control of sequential behavior in oscillatory EEG activity. *eLife*, *7*.

Kikumoto, A., & Mayr, U. (2017). The nature of task set representations in working memory. *Journal of Cognitive Neuroscience*, *29*(11), 1950–1961.

Kikumoto, A., Hubbard, J., & Mayr, U. (2015). Dynamics of task-set carry-over: evidence from eye-movement analyses. *Psychonomic bulletin & review*, 1-8.

Mayr, U., Kleffner-Canucci, K., Kikumoto, A., & Redford, M.A. (2014), Control of task sequences: What is the role of language? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 40(2).

ACKNOWLEDGMENTS

I wish to express sincere appreciation to Dr. Ulrich Mayr for his guidance over the years since I joined the lab as an undergraduate research assistant, and for patiently working together to embrace new challenges. Additionally, I would like to thank Dr. Brice Kuhl, Dr. Sarah Dubrow, and Dr. Ian Greenhouse for their thoughtful inputs on this manuscript. I am indebted present and former research assistants in the Cognitive Dynamics Lab who have provided support: Lauren Williams, Selina Robson, Chihoko Hayashi, Caitlin Corona, Katelyn Occhipinti, Joshua Karpf, Aran Lenart, Dagger Anderson, Isabella Dickerson, Tesufuaishin Sameshima, Min Zhang, Megan Carson, Ali Byers, Jena Kunimune, Vy Tran, Christian Manalansan, Jiafan Jia, Emily Stutz, Chelsea Roertson, and Anastasia Browning, and present and former colleagues in the lab: Jason Hubbard, Pablo Morales, Melissa Moss, Theo Schäfer, Samuel Lippl, and Kerstin Froeber.

Dedicated to my wife Ayaka Kikumoto and my son Tasuku Kikumoto who provided support (and data), and my parents Yonai Kikumoto, Masayo Kikumoto, Ichiro Yuasa, and Akemi Yuasa, and my friends Dave Schenderlein, Beth Schenderlein and Sarah Schenderlein.

TABLE OF CONTENTS

Chapter Page

LIST OF FIGURES

Figure Page

Figure Page

LIST OF TABLES

CHAPTER I.

INTRODUCTION

Every situation we experience is unique, and yet we are able to usually act appropriately by applying relevant knowledge or rules that is useful across different events. For example, we can solve new jigsaw puzzles that we have never touched before or drive a rented car in a foreign land or by applying the general knowledge associated with the task (e.g., "find all the edge pieces" or "slow down at the corner"). Such flexible, goal-directed actions rely on abstract, generalizable rules that can configure behavior to a range of specific circumstances. Even for a simple action such as joining puzzle pieces together, various task-relevant features—the shape and color of pieces, the location of missing puzzles, as well as abstract rules and strategies—need to be adequately represented and integrated to accomplish the overarching goal. Thus, action selection entails cognitive control processes that enable us to adapt perceptual and response selection processes to guide thoughts and actions in accordance with internal goals[1–7].

In task spaces that are relatively consistent over time, it would be adaptive to acquire or use action rules that summarize possible stimulus-response functions, discarding non-critical aspect of the task that are specific to situations. For instance, for the rule of "finding all the edge pieces" to be effective for many puzzles, the irrelevant properties of the task (e.g., the shape of pieces) should be detached from the action rule. Past studies have indicated humans can learn[8–10] or be instructed[11–13] abstract action rules or task sets to link external inputs to specific responses (i.e., stimulus-response mapping). Though, critically, because abstract representations are detached from the specific contexts, action rules need to be "translated" to the specific conditions for the current goal. Yet, surprisingly, we currently know little about how exactly abstract action rules

are adapted towards specific sensory and/or motor requirements, as goal-directed actions are planned and executed. Such translation processes determine how abstract action rules become usable to enable specific actions, thus efficiency in configuration of action rules could undermine their values.

One class of dominant theories of action control suggests that action rules prompt specific stimulus and/or response settings in a strictly feedforward and stage-specific manner[14,15]. Here, action selection is considered as a hierarchically nested process, where the higher-level action rule representation constrain selection at the subordinate level [10,16–20]. For example, Kleinsorge and Heuer (1999) proposed a model to explain the pattern of priming costs during action transition that further incorporates an "update" signal propagating down from the higher to lower hierarchical levels in a unidirectional manner[16–20]. Critically, in these models, hierarchical selection recruits level-specific action representations in a relatively independent manner, thus action rules remain functionally separable from other relevant features. These notions are consistent with findings indicating localized cortical functions with hierarchical structures[16,21–23].

An alternative view is that action selection requires a common representational space in which task-relevant features (e.g., sensory, motor, and even abstract action rules) are combined into highly integrated, cross-modal, conjunctive representations, sometimes referred to as *event files* or *task files*[24–28]. Essentially, conjunctive representations instantaneously store the entire profile of action control that enables specific actions[29,30]: a set of bindings that temporarily connect representations of the relevant or salient features of the perceptual event, the accompanying action, and the task setting. This class of theories suggest that the integration of task-relevant features is the critical precursor of successful action control. Consistent with such a view, recent evidence in non-human

primates studies indicated that a majority of neurons are tuned to the mixture of task features in a nonlinear manner, encoding the conjunctive information of goal-relevant features that produce unique functional and computational properties[31–35].

Behavioral studies in humans have provided some evidence supporting event-file theory. For instance, the theory predicts that during the lifetime of an event-file for specific actions, an encounter with one (or more) of the bounded features causes the automatic retrieval of the whole event-file. This in turn produces a characteristic, partial-overlap priming pattern (further discussed in Chapter II), which has been considered as the key indicator of event files. Mayr and Bryck (2005) showed that also abstract action rules can become part of event files. In the same vein, several studies have reported the interaction of repetition priming effect (e.g., response-repetition) with other task parameters[36–38], which could be parsimoniously explained by binding process of action features. However, because these priming effects occur as an *aftereffect* of event-file formation, they alone provide no information about how the conjunctive representations emerge *during* action selection. Because there has been no viable method to directly track the formation of nonlinear, conjunctive representations in humans, which is expected to occur at a subsecond scale, it has been unclear how action rules representations make contract to other representations of sensory and/or response features during action selection.

To this end, in the current work, we applied a time-resolved, decoding approach to EEG while participants performed a rule-based action selection task[26] by capitalizing on the multivariate analysis techniques recently applied to the diverse neural activity in humans[39–43]. This allowed us to directly track multiple action representations, including potential conjunctions that integrate action features in a nonlinear manner. Our method

allowed us to flexibly establish multiple action-relevant dimensions without explicitly demanding subjects to bind action features. It further allowed us to characterize dynamic control of task representations of both basic (e.g., stimuli, responses and action rules) and conjunctive features in a temporally precise manner, and to relate the quality of these representations to the theory-relevant behavioral outcomes on the level of single trials. Overall, in a series of studies, we found robust evidence supporting that the conjunctive representations are a critical driver of goal-directed actions.

CHAPTER II.

ACTION SELECTION AND TRANSITION

Introduction

Flexible, goal-directed action requires the use of abstract rules that can be applied to a range of specific situations. However, we know little about how abstract rule representations connect with lower-level sensory or response representations, as a specific action is planned and executed. In traditional stage-based processing models, information flows from sensory to response in a cascade of relatively independent representations[1-5], that are specified by the relevant action rule[6]. An alternative view is the idea of a common representational space in which all action-relevant features (e.g., sensory, motor, and even abstract action rules) are combined into highly integrated, conjunctive representations, sometimes referred to as *event files* or *task files*[7-10]. By tying all relevant features together into a common, integrated representation, a specific action becomes executable. Therefore, these representations are a critical condition for successful action control and selection.

Once formed, however, an event file can also get in the way of subsequent actions, as indicated by a characteristic pattern of priming effects[11]. Specifically, when consecutive trials require event files that share either all or none of the constituent features, actions are executed relatively fast. However, when only some, but not all features overlap across trials, then response-times or errors increase, a pattern that event-file theory explains as the cost of "unbinding" the overlapping features from the no-longer needed event file. Such partial-overlap costs emerge even when complete S-R associations repeat across trials while the abstract rule changes, indicating that just like any sensory or response features, rules can become part of event files[10].

The partial-overlap pattern is currently the key empirical indicator of event files. However, because it is an *aftereffect* of event-file formation, this pattern provides no information about how conjunctive representations behave during response selection and whether or not they are indeed a critical precursor of successful action. Moreover, partial-overlap costs can also be explained by alternative models that do not assume integration between different codes during action selection. For example, Kleinsorge & Heuer[12, 13] proposed a strict hierarchical separation between the level of rules and the level of stimulus/response selection. The pattern of partial-overlap costs arises from the assumption that a switch of action codes on the highest level (i.e., rules) propagates down to the lower levels (i.e., stimulus-response codes). As a result, when only the rule changes, but the response stays constant, the now inappropriate lower-level specification will have to be reverted, leading to performance costs.

To test the event file model against accounts that do not assume integration of action-relevant features as a critical step during action selection, it is important to directly track the multiple representations that could concurrently become active during action selection—including potential conjunctions between stimuli, responses, and even action rules. In the current study, we used the EEG signal to decode information about action-relevant representations in a time-resolved manner[14-16] via representational similarity analysis (RSA)[17-19] as participants selected responses to location stimuli on the basis of randomly cued, spatial transformation rules (Fig.1ab)[10]. Experiment 1 allowed us to decode conjunctions that were specific for particular rules, but without differentiating between S-R conjunctions and conjunctions that also integrated abstract rules (i.e., rule-S-R conjunctions). In Experiment 2, we replicated all major results from Experiment 1, but also used an expanded task space that allowed us to test whether or not abstract rules

can become integrated into conjunction representations. Across both experiments, we found strong evidence for conjunctive representations—including rule-S-R conjunction in Experiment 2. Consistent with predictions from event-file theory, conjunctions were robust and unique predictors of variability in performance, and were related to the pattern of partial-overlap priming costs.

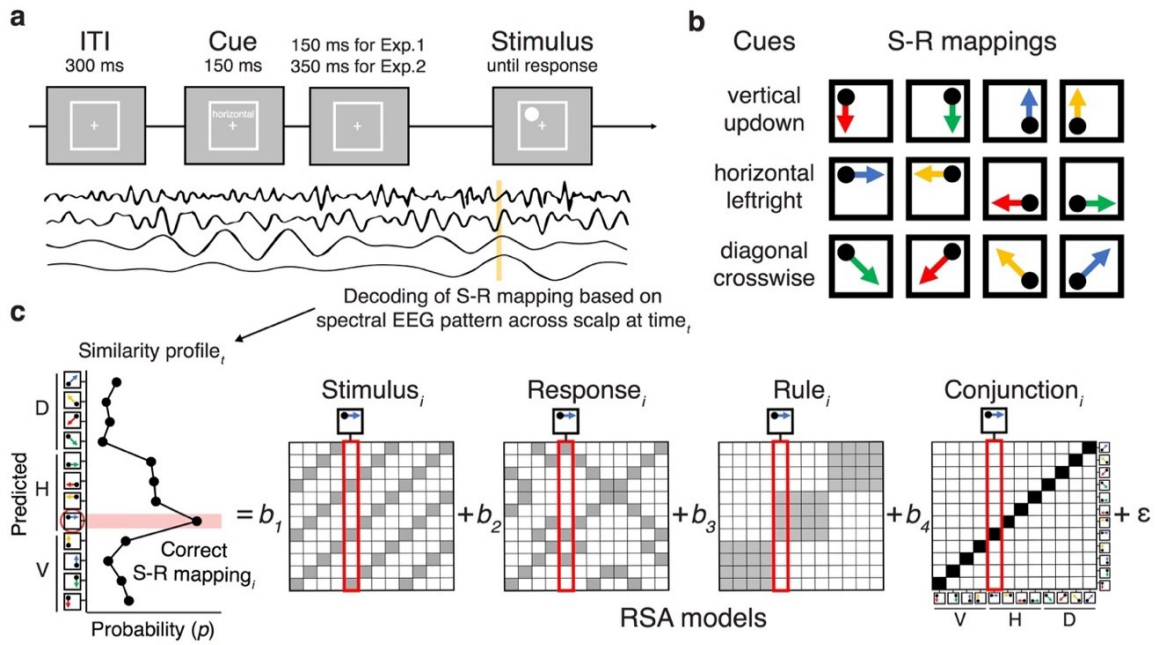


Fig. 1. a, Sequence of trial events in the rule-selection task for both Experiment 1 and 2. **b**, Spatial translation rules mapping specific stimuli to responses in Experiment 1. Two different cue words were used for each rule. **c**, Schematic steps of the representational similarity analysis. The 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 idealized profile of classification probabilities shows an example 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.

Results

Experiment 1

Behavior

For all analyses, error-trials, post-error trials, and trials in which RTs were larger than 99.5 percentile of the RT distribution were excluded. Consistent with previous work[10], we observed partial-overlap costs in RTs and errors as a function of the different trial-to-trial transitions (Fig. 2): When the rule, the stimulus, and thus also the response repeated or when all changed, responses were fast and accurate, whereas costs emerged in the case of partial updates of either rules or stimuli/responses (for statistical analysis, see Supplementary Table 1).

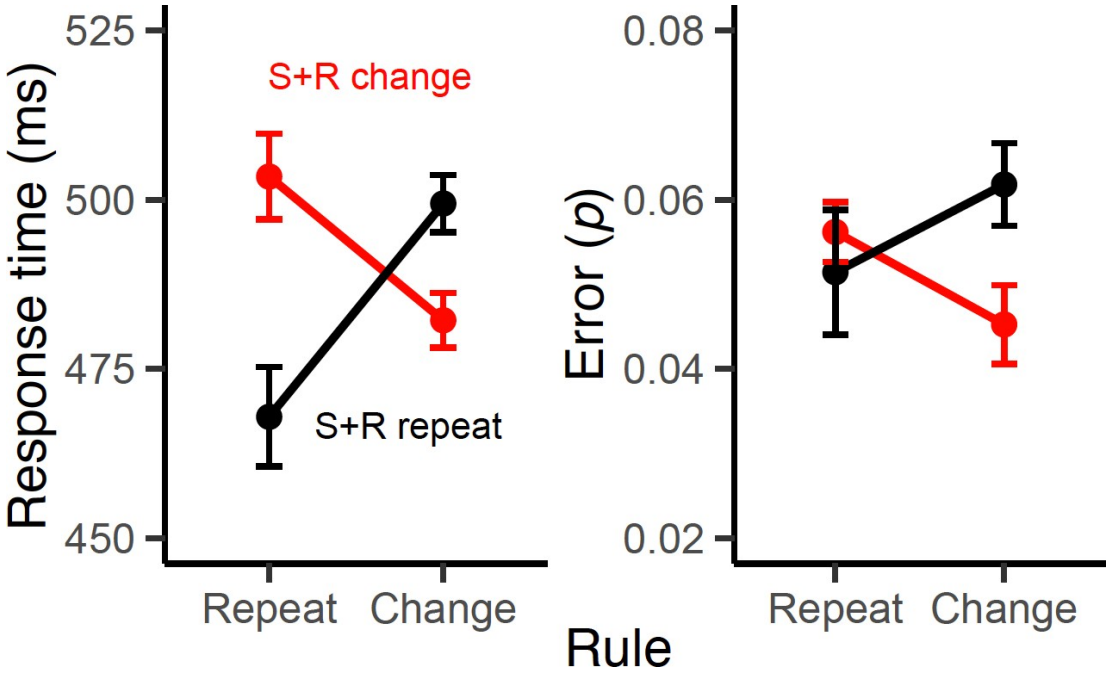


Fig. 2. Mean response times (RTs) and errors for Experiment 1 as a function of rule repetition/change factor and the stimulus-response repetition/change factor. Error bars specify 95% within-subject confidence intervals.

Tracking Representational Dynamics

While the pattern of RTs and errors is consistent with predictions from the event-file model, by itself it is not sufficient to draw strong inferences about the role of

conjunctive representations during action selection. Fig. 3a shows the results of the time-resolved RSA performed on the level of single trials. Consistent with previous results, the cascade of decoded representations unfolds in a manner that is consistent with the expected flow of information: The rule is activated during the pre-stimulus phase, followed by a strong expression of the stimulus, and finally by the response[14,16]. Critically, the conjunctive representation can be decoded during the entire post-stimulus period (Fig. 3a), and clearly peaks before response representations fully develops (Supplementary Fig. 7). These effects were significant even though we accounted for subject-specific differences in RTs between action constellations and therefore cannot be explained in terms of unspecific difficulty differences between action constellations (Supplementary Fig. 1).

To test the prediction from event-file theory that conjunction representations are critical for action selection, we regressed trial-to-trial variation in RTs onto the strength of each expressed representation. Using multilevel modeling, we performed these analyses for each time-point and with all predictors entered simultaneously. The resulting "impact-trajectories" are shown in Fig. 3b; statistical results for a-priori selected time intervals are summarized in Table 1. Note that negative *t*-values indicating that stronger representations lead to faster responding. Consistent with the prediction from event-file theory, the conjunctive representation was the most dominant predictor of performance. Combined, these results indicate that conjunctive representations emerge during response-selection, concurrently with the representations of constituent features, and predict upcoming behavior over and above the influence of the constituent representations.

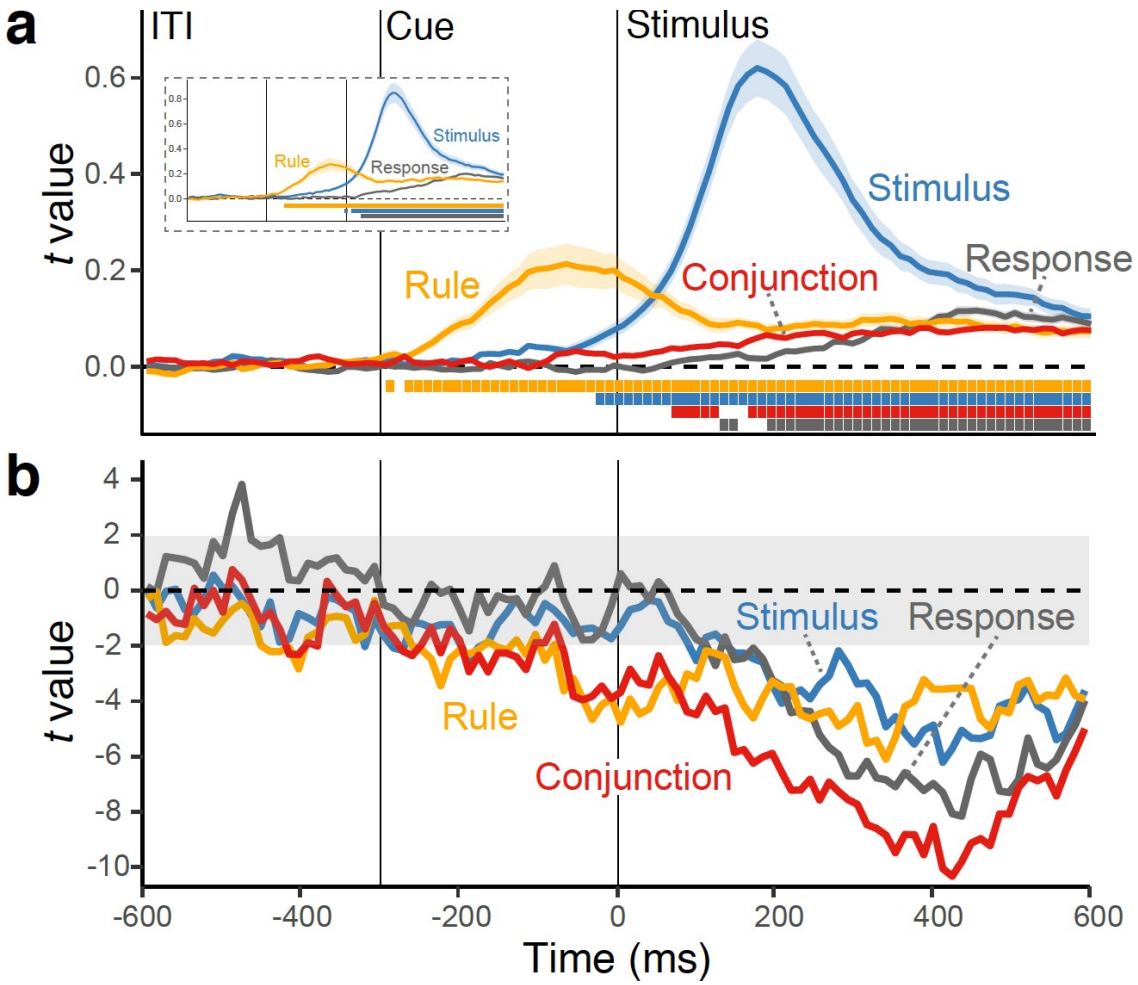


Fig 3. a, Average, single-trial *t*-values associated with each of the basic features and their conjunction derived from the RSA analysis (see Fig. 1c). Shaded regions specify the standard error around the mean. The colored squares at the bottom of the figure denote the significant time points using a non-parametric permutation test. The insert shows RSA fit scores when the conjunction was not included as predictor in the analysis. **b**, Time-course of *t* values from multilevel, linear models predicting the variability in trial-to-trial RTs (the "impact" of representations on behavior), using RSA scores of all features as simultaneous predictors.

Table 1.
Predicting trial-by-trial RTs using the strength of decoded representations for Exp. 1

	Pre-stimulus		Early Post-stimulus		Late Post-stimulus	
Variable	*b* (*se*)	*t*	*b* (*se*)	*t*	*b* (*se*)	*t*
Rule	-.021 (.005)	-3.89	-.004 (.007)	-5.71	-.037 (.007)	-4.97
Conjunction			-.061 (.008)	-7.23	-.093 (.011)	-8.07
Stimulus			-.014 (.008)	-1.82	-.041 (.012)	-3.49
Response			-.027 (.006)	-4.12	-.068 (.012)	-5.81

Conjunctive Representations and Partial-Overlap Costs

In order to directly connect the EEG-decoded conjunctive representations with event-files, we examined whether and how these representations relate to the partial-overlap priming pattern. As Fig. 4a shows, the strength of decoded conjunctions expresses the partial-overlap pattern. Conjunctive representations were particularly strong exactly in those transitions in which RTs were fast (i.e., when either everything repeats or everything changes, see Fig. 2). Conjunctive representations showed the partial-overlap pattern in the correct direction during early post-stimulus phase, b=-.024, SE=.010, $t(20)$=-2.58, but not in the late post-stimulus phase, b=-.004, SE=.010, $t(20)$=-.39, and none of the constituent features showed the critical interaction pattern, all $t(20)$>-.21.

Another important prediction that can be derived from the event-file model is that strong conjunctions should be particularly difficult to "unbind" on the following trial. Thus, the strength of conjunctions on trial n-1 should predict partial-overlap costs on trial n. Our results, shown in Fig. 4b, confirm this prediction: A stronger conjunctive representation in n-1 trial, late in the selection period led to a greater RT partial-overlap costs on the next trial, b=-.025, SE=.011, $t(20)$=-2.25. Again, this pattern was unique for conjunction representations. None of the constituent features had comparable effects on next-trial performance, all $t(20)$>-.05. Taken together, the behavior of decoded conjunctive representations is highly consistent with predictions from the event-file model.

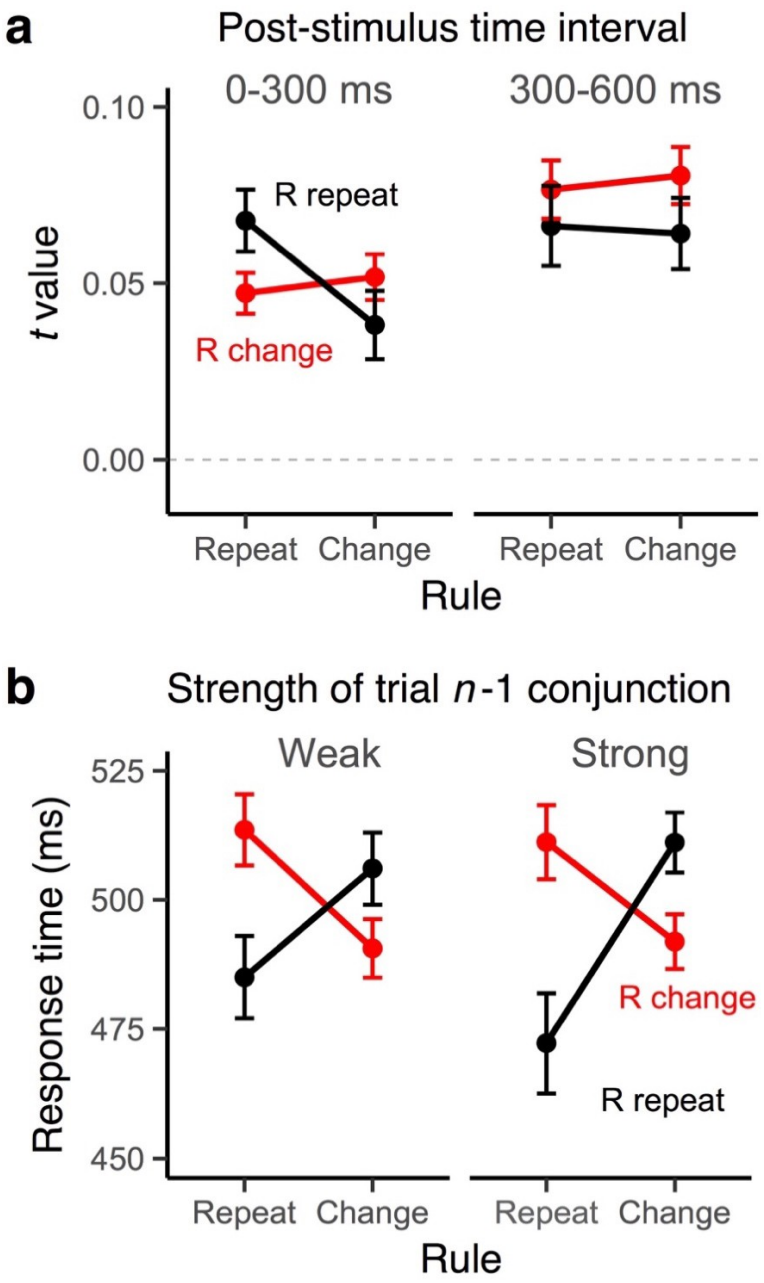


Fig. 4. a, Average RSA scores of the conjunction model as a function of rule repetition/change and the stimulus-response repetition/change factors for two the early (0-300 ms) and the late (300-600 ms) periods in the post-stimulus interval **b**, Modulation of partial-overlap costs on RTs in trial *n* as a function of the strength of conjunction codes (median split) in trial *n* -1. Error bars specify 95% within-subject confidence intervals.

Experiment 2

The results in Experiment 1 suggest that action selection recruits conjunctive representations and that the strength of these representations is predictive of trial-to-trial

performance, as postulated by the event file perspective[8, 11, 14]. Yet, because each action rule specified a unique set of S-R links, the observed conjunctive representations could consist of any combinations of the rule and/or stimulus/response features—leaving it ambiguous to what degree abstract rules were integrated into conjunctive representations. Thus, in Experiment 2, we attempted to tease apart two different types of conjunctions: (1) rule-independent conjunctions between stimuli and responses (S-R conjunctions) and (2) rule-specific (rule-S-R) conjunctions that integrate abstract action rules with S-R links. To this end, we introduced four action rules (i.e., vertical, horizontal, clockwise, and a counterclockwise; Fig. 5a), which allowed S-R conjunctions that shared the same S-R links, but different abstract rules (e.g., a dot at the top-left corner requires a bottom-left response using either the vertical rule or the counterclockwise rule)[10]. The inclusion of such same-S-R pairs allows dissociating conjunctions that did integrate rules (rule-S-R) from rule-unspecific conjunctions (S-R).

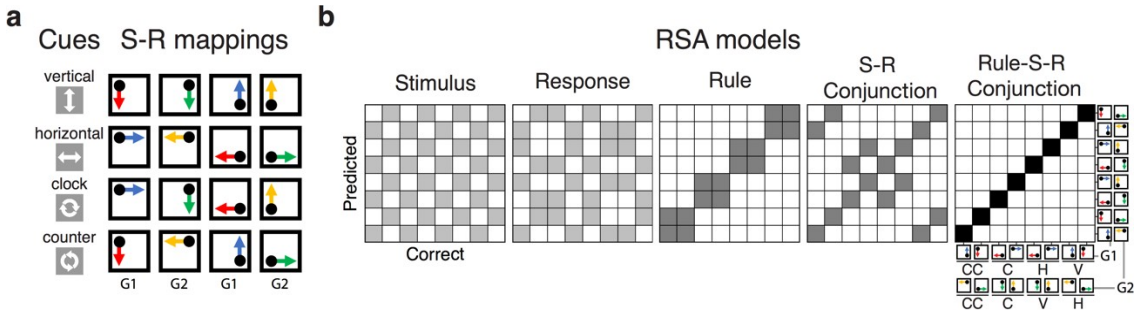


Fig. 5. **a**, Spatial translation rules mapping specific stimuli to responses in Experiment 2. Either words or symbols were used as cues for each rule. S-R mappings were divided into two groups: G1 group (cases where a dot appeared at the top-left or bottom-right corner) and G2 group (cases where a dot appeared at the top-right or bottom-left corner) for the decoding analysis. **b**, Models for representational similarity analysis (RSA) in Experiment 2. The S-R conjunction model assumes a similar pattern for the specific combination of the stimulus and response irrespective of rules. The rule-S-R conjunction model expects a unique pattern for the configuration of each rule/stimulus/response combination. To completely orthogonalize action features, RSA was performed separately for G1 and G2 subsets of S-R mappings, requiring identical 8 x 8 model matrixes for each group. Analyses were performed separately within each subset and coefficients were averaged within subjects.

Behavior

The same trial-exclusion criteria as in Experiment 1 were used for all analyses in Experiment 2. We replicate the partial-overlap costs on RTs and errors from Experiment 1 and a previous report using the same paradigm[10] (Fig. 6): Critically, repetition of rule-S-SR settings produced RT and error benefits, whereas any partial updates (including S-R repetitions) generated costs. We focused on trials in which both stimulus and response-features covary (i.e., complete S-R changes or repeats) as a function of switching of rules because they provide a direct test for potential differences in the rule-specific and rule-independent conjunctions (for statistical analysis, see Supplementary Table 2).

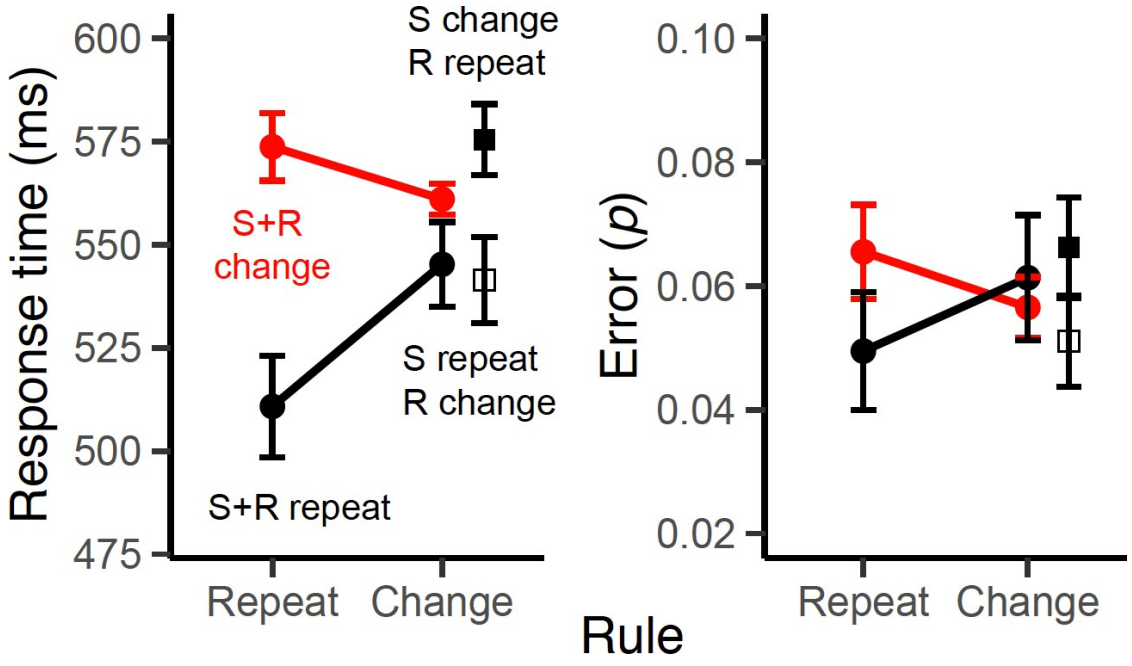


Fig. 6. Mean RTs and errors for Experiments 2 as a function of the rule repetition/change, stimulus repetition/change, and the response repetition/change factors. Note that, in rule-repeat trials, partial updates of S-R settings (e.g., S change + R repeat) are not possible. Error bars specify 95% within-subject confidence intervals.

Decoupling Rule-specific and Rule-independent Conjunctions

We used the same analysis approach as in Experiment 1, only that here we included RSA models for both rule-specific (rule S-R) conjunctions and rule-independent

(S-R) conjunctions. We found evidence that both types of conjunctions emerged over and above the constituent features (Fig. 7a). The rule-S-R conjunctions became active right after stimulus onset and were sustained robustly during the selection period. Again, activation of these representations preceded the emergence of response information (Supplementary Fig. 7). In contrast, the rule-independent, S-R conjunctions appeared immediately after rule-S-R conjunctions, but remained relatively weak compared to other action representations. We also replicated the pattern of results from Experiment 1 for the constituent features, with the exception that the rule representations diminished after stimulus onset (Fig. 7a). Excluding conjunction models restored the post-stimulus rule representation, suggesting that the rule S-R conjunction model captures the same variance as the rule model explains in this phase of action selection (Fig. 7a inset).

Table 2.
Predicting trial-by-trial RTs using the strength of decoded representations for Exp. 2

	Pre-stimulus		Early Post-stimulus		Late Post-stimulus	
Decoded Variable	*b* (*se*)	*t*	*b* (*se*)	*t*	*b* (*se*)	*t*
Rule	-.019 (.005)	-3.80	-.017 (.013)	-1.34	-.007 (.003)	-2.54
Rule S-R Conj.			-.061 (.012)	-5.00	-.019 (.002)	-8.70
S-R Conj.			-.053 (.014)	-3.87	-.022 (.003)	-7.32
Stimulus			-.010 (.007)	-1.42	-.008 (.002)	-3.94
Response			-.037 (.014)	-2.57	-.016 (.002)	-6.60

Next, we examined again, which representations were the main driver of action selection. As shown in Fig. 7b, both rue-S-R conjunctions and S-R conjunctions explained substantial, and independent variability in trial-to-trial RTs, over and above the variance explained by the constituent features (Table 2 for the statistical results and Supplementary Fig. 4 for results from standard decoding analyses). These results replicate the findings from Experiment 1 that conjunctions are indeed critical of efficient action selection. In addition, they clarify that both rule-independent and rule-specific

conjunctions are about equally important in predicting behavior, with possibly a slight edge for the rule-specific conjunctions.

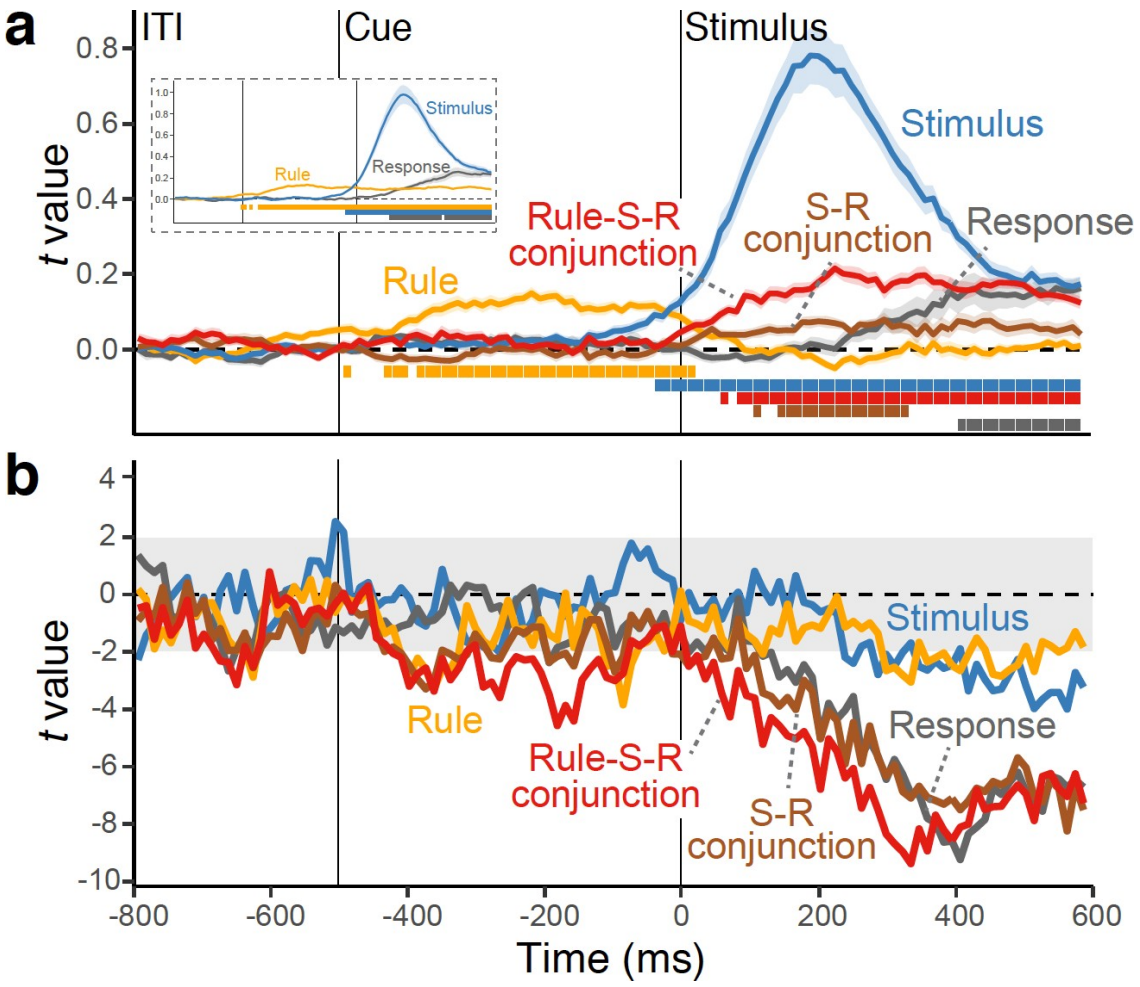


Fig. 7. **a**, Average, single-trial *t*-values associated with each of the basic features and their conjunctions, derived from the RSA analysis (see Fig. 1c and 5). Shaded regions show the standard error around the mean. The colored squares at the bottom of the figure denote the significant time points using a non-parametric permutation test. The insert shows the same RSA fit scores when the conjunctions (i.e., rule S-R conjunction model and S-R conjunction model) were not included as predictors in the RSA analysis. **b**, Time-course of *t* values from multilevel, linear models predicting the variability in trial-to-trial RTs (the "impact" of representations on behavior), using RSA scores of all features as the simultaneous predictors. RSA model vectors for stimulus, response, rule, and conjunction representations. RSAs were performed separately within a subset of action constellations (i.e., G1 and G2) to orthogonalize all features (see the Method and Fig. 5ab for details).

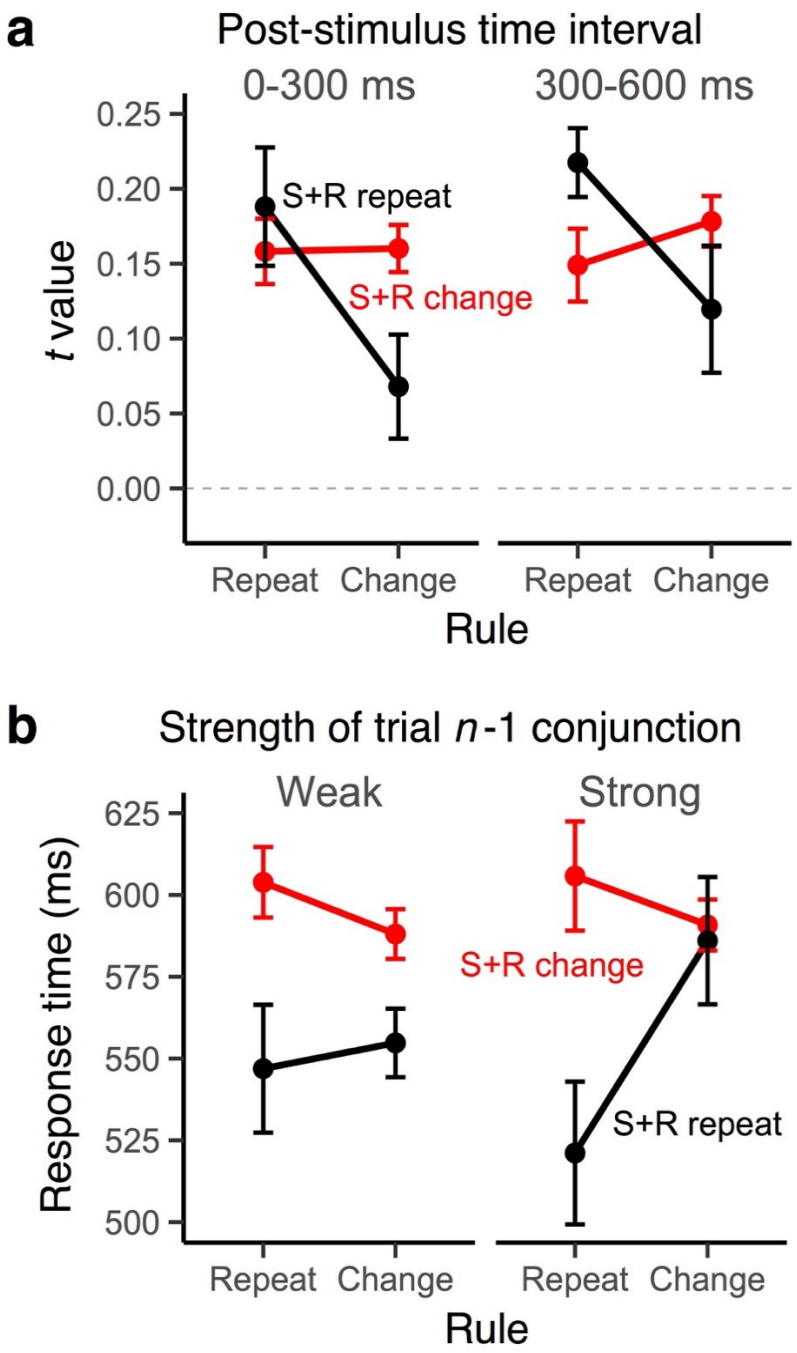


Fig. 8. **a**, Average RSA scores of the rule-S-R conjunction as a function of the rule repetition/change and the stimulus-response repetition/change factors for early (0-300 ms) and late (300-600 ms) periods in the post-stimulus interval. **b**, Modulation of partial-overlap priming patterns in trial *n* as a function of the strength of the rule S-R conjunction in trial *n* -1 (median split). Error bars specify 95% within-subject confidence intervals.

Conjunctive Representations and Partial-Overlap Costs

As in Experiment 1, we also examined the relationship between conjunction codes and behavioral partial-overlap costs. We found that only the strength of rule-S-R

conjunctions showed the partial-overlap costs, b=-.021, SE=.009, t(21)=-2.22, for the early selection phase; b=-.021, SE=.009, t(21)=-2.24, for the late selection phase (Fig. 8a). None of the constituent features, t(21)>-.72), or S-R conjunctions showed such an effect, b=-.012, SE=.009, t(21)=1.27 for the early selection phase; b=-.007, SE=.010, t(21)=-.72, for the late selection phase. In addition, the strength of late rule-S-R conjunctions on the previous trial again significantly modulated RT partial-overlap costs on the next trail (Fig. 8b), b=.031, SE=.011, t(20)=2.81. This pattern was absent for constituent features, all t(21)<.38, and S-R conjunctions, b=-.009, SE=.011, t(21)=-.85. Thus, only the conjunctions that integrate rule information show a tight relation with the main behavioral indicator of event files, the partial-overlap cost.

Discussion

We tested whether integrated, conjunctive representations between task-relevant features emerge during action selection, as postulated by event-file theory[7, 8]. In our paradigm action settings had to be updated flexibly for each trial, creating unique constellations between rules, stimuli and responses. We combined a standard, linear decoding approach with a subsequent, time-resolved RSA in order to track the emergence of conjunctive representations and their constituent features over time, and for each individual trial. In Experiment 1, conjunctions could entail any pairwise, or complete combination of rule, stimulus, or response features; in Experiment 2, we were further able to dissociate between rule-S-R conjunctions, and rule-independent S-R conjunctions.

The time course of decoded information showed a highly plausible cascade of action representations (rule, stimulus, and then response), and most critically, we found robust evidence for conjunctive representations—emerging shortly after stimulus onset and then persisting until response execution. Analyses with response-locked EEG data

fully confirmed this pattern of results (Supplementary Fig. 7). The fact that conjunctive representations are continuously present from stimulus processing to response execution is consistent with their role in translating perceptual codes into response codes based on the current task rules. Even though the strength of conjunctive representations was on average much weaker than that of the constituent features, they were highly robust and consistent within individuals (Fig. 3 and Fig. 7 and Supplementary Fig. 9). Even more importantly, conjunctive representations were strong and unique predictors of trial-by-trial variability in RTs, over and above other constituent features. These results are difficult to reconcile with traditional stage theories[1-5], where information flows in a strictly feed-forward manner and therefore does not allow the emergence of integrated representations. These results are also inconsistent with hierarchical control models that assume independent selection processes on different hierarchical levels[12, 13]. Instead, our results indicate that action selection is established by tying together disparate, task-relevant features from the entire selection event into a common representation.

The fact that in Experiment 2 rule-S-R conjunctions *and* rule-independent, S-R conjunctions emerged is an important result in its own right. It suggests that integrated representations that match the contingencies in the environment can develop in parallel, and on different levels of abstraction. This combination of both highly specific and rule-general representations can account for the fact that S-R associations learned within one rule can transfer to another rule, albeit in a limited manner[10, 21]. It is also consistent with the proposal that event files themselves can possess an internal, hierarchical organization[9].

A key behavioral indicator of event files is the partial-overlap priming pattern, which entails benefits when all action features either repeat or change and costs when there is partial overlap of features across trials[10, 11]. In both experiments, we found that

this pattern not only in RTs and errors (Fig. 2 and Fig. 6), but also in the strength of conjunctions (Fig. 4a and Fig. 8a). Even more importantly, the strength of conjunctions on trial *n*, predicts the size of partial overlap costs on trial *n*+1 (Fig. 4b and Fig.8b), suggesting the stronger action features are tied together into conjunctions, the harder it is to "unbind" them on the following trial in order to integrate them into a new conjunction. Recent behavioral studies have raised questions about whether the strength of partial-overlap costs is explained by the strength of the initial binding, or instead by difficulty of selectively retrieving integrated, action-relevant features[8]. While the present results do not rule out the contribution of retrieval-related effects, they do point to the "binding strength" of the original conjunction as a critical factor that determines partial-overlap costs.

It is particularly important that the tight relationship with the partial-overlap pattern was only found for conjunctions (i.e., rule-S-R conjunction in Experiment 2), thus functionally dissociating conjunctions from their constituent codes. Moreover, the results in Experiment 2 also indicated that only rule-specific S-R conjunctions were related to the partial-overlap cost, not however the rule-independent S-R conjunctions. As noted, for Experiment 1, the task design did not allow firm conclusions about whether or not conjunctions contained rule-specific information. However, the conjunctions in Experiment 1 showed a similar priming pattern as the rule-S-R conjunctions in Experiment 2, suggesting integration of not just stimuli and responses, but also of rules in both experiments.

While our results indicate with high resolution when representations of specific features and feature combinations are activated, they provide no neuroanatomical information (see Supplementary Results). Cell-physiological work with monkeys and

human, neuroimaging work indicates that the representation of task-relevant features, including rules, is distributed across large areas frontal and parietal cortex[22, 23]. From animal models, there is substantial evidence that the hippocampus and the frontal cortex are particularly important for representing conjunctive information[24, 25]. Human neuroimaging work also mainly implicates the hippocampus [26, 27]; attempts to decode task representations in the frontal areas have proven more challenging[28], but have also seen some recent success[29, 30].

In nonhuman primates, single cell recordings have also shown that while basic task features (cues, rules, stimuli, and responses) are encoded across various frontal and parietal areas during rule-based action selection[22, 31], a substantial proportion of recorded neurons are tuned to the mixture of multiple features in a non-linear manner[24, 32]. Such heterogeneous, neural responses allow both efficient, linear read-out of information to downstream neurons and can also code high-dimensional, conjunctive information[33].

An important finding from this research is that the degree of conjunctive information coded in recorded neurons is functionally distinct from the representation of linear features. For example, high-dimensional, non-linear information was found to be highly robust on correct trial, but was largely missing on error trials, whereas low-dimensional information is equally strong on correct and error trials[24]. This pattern is consistent with our finding that the strength of conjunctive representations uniquely predicts trial-by-trial performance, beyond the predictive strength of constituent, simple features (Fig. 3b and Fig. 7b). Further evidence for a functional dissociation comes the finding that conjunctive representations express trial-to-trial transitions (i.e., the partial-overlap priming pattern) in a qualitatively different manner than the constituent feature representations (see Fig. 4 and Fig. 8).

These results about the relevance of conjunctions for efficient action selection and the mismatch priming pattern also directly confirm predictions from event-file theory. Therefore, they provide an important, missing link between two, so-far distinct lines of research: The relatively abstract, event-file theory, designed to explain the architecture of human action selection, and the recent advances from animal research about the neural implementation of high-dimensional, non-linear representations. Beyond the current demonstration of the role of conjunctive representation in human action control, there is a range of important, open questions. For example, we do not know how these representations are constrained by capacity limitations[34], to what degree they allow integration of action outcomes or goals[35], or how they change through experience (see Supplemental Fig. 9)[10]. The decoding approach used here, promises answers to these and related questions.

Method

Participants

A total of 44 people participated after signing informed consent following the protocol approved by the University of Oregon's Human Subjects Committee in exchange for the compensation of $10 per hour and the additional performance-based incentive. Participants with excessive amount of EEG artifacts (more than 35% of trials) were removed from further analysis. As a result, we retained 20 out of 22 participants for Experiment 1 and 21 out of 22 for Experiment 2.

Stimuli, Tasks and Procedure

Participants performed a cued rule-selection task, in which one of the pre-instructed action rules, on trial-by-trial basis, was randomly selected to determine possible S-R mappings[10]; Fig. 1b). 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. Each action rule specified four S-R mappings using a simple spatial transformation rule. For instance, the vertical rule mapped the left-top circle to the bottom-left response as a correct response and vice versa. We used two cues for each rule (a pair of verbal cues in Experiment 1 and symbol/word pair in Experiment 2) that appeared in either even or odd trials to prevent immediate cue repetitions. Thus, cues, rules, responses, and stimuli were orthogonalized, and the combination of these features generated unique action constellations.

In Experiment 1,"vertical", "horizontal" and "diagonal" rules were randomly cued (i.e., 66.6% switch rate). In Experiment 2, "vertical", "horizontal", "clockwise" and "counterclockwise" rules were used (i.e., 75% switch rate; Fig. 2c). Here, half of S-R links were shared across rules (e.g., a left-top circle leads to a left-bottom response in both the vertical and the counterclockwise rule). This allowed us to generate transitions between trials with rule changes but repetitions of S-R links (Fig. 1c).

There were two practice blocks and 200 experimental blocks in both studies. Participants were instructed to respond as fast and accurately as possible to complete as many trials as possible within each 16-second block. Trials that began within the 16 seconds were allowed to complete. Participants were given a performance-based incentive for trials with RTs faster than the 75th percentile of correct responses in the preceding blocks when 1) the overall accuracy was above 90 percent and 2) there were more than 7 completed trials in a given block. While performing the task, participants were asked to rest the index finger of their dominant hand in the center of the four keys in matrix and to hit the correct key. All stimuli were created in Matlab (Mathworks) using

the Psychophysics Toolbox [36, 37] and were presented on a 17-inch CRT monitor (refresh rate: 60 Hz) at a viewing distance of 100 cm.

EEG recordings and preprocessing

EEG data was first epoched by 18 second intervals to include all trials within a block (see the Appendix A. EEG RECORDING AND PREPROCESSING). After time-frequency decomposition was performed (see the Appendix B. TIME-FREQUENCY ANALYSIS), these epochs were further segmented into trial-to-trial epochs (-600 ms to 600 ms intervals for Experiment 1 and -800 ms and 600 ms intervals for Experiment 2, relative to the onset of a stimulus). 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. For all EEG analyses, error trials, post-error trials and trials with exceedingly slow RTs (i.e., slower than 99.5% of all responses) were excluded to be consistent with behavioral analyses.

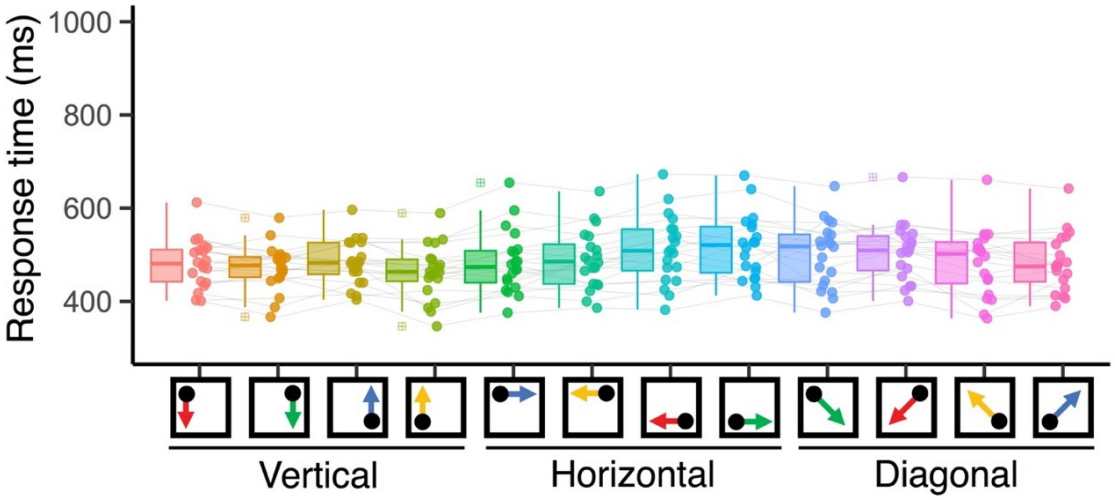
Representational Similarity Analysis

Our goal was to obtain information about the strength of each feature and conjunction on the level of individual trials and timepoints within trials. This required a two-step procedure. First, we performed a linear decoding analysis to discriminate between all 12 different action constellations in Experiment 1, or 16 constellations in Experiment 2. Specifically, we performed a penalized linear discriminant analysis using the caret package in R[39-41]. At every time sample point, the 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

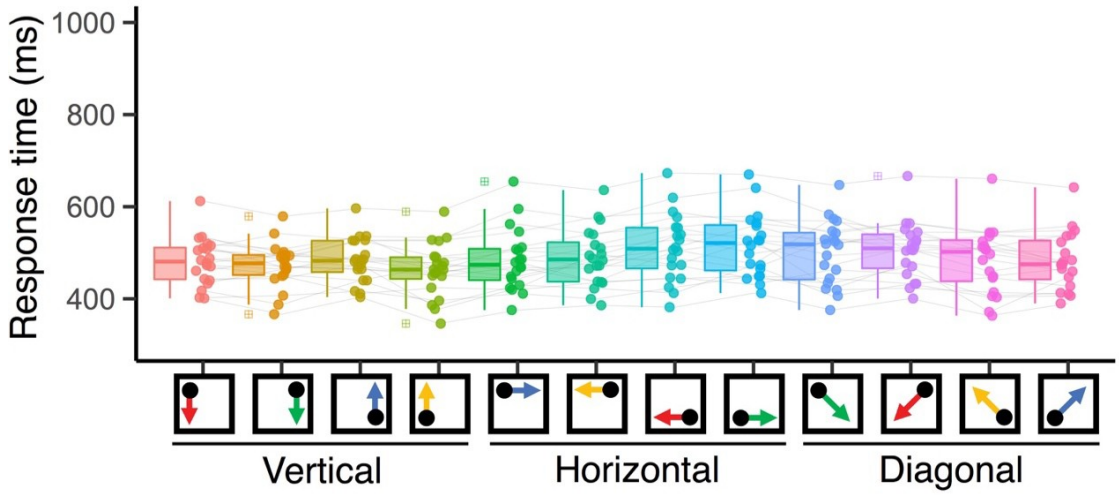
electrodes) to train decoders. Within individuals, these data points were z-transformed across electrodes at every sample to remove the effects that uniformly influenced all electrodes. We used a *k*-fold repeated cross-validation procedure to evaluate the decoding results[42], by randomly partitioning single-trial EEG data into four independent folds. 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 parameters. 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 RSAs[17] to each profile of classification probabilities in order to determine their underlying similarity structure for each time point and trial. Specifically, we regressed the confusion vector onto model vectors as predictors, which were derived from a set of representational similarity model matrixes. Each model matrix uniquely represents a potential, underlying representation (e.g., rules, stimuli, responses and conjunctions; Fig. 1c and Fig. 5b). 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. Thus, we obtained coefficients for each of the four model vectors (e.g., rule, stimulus, response, conjunction for Experiment 1). These coefficients (i.e., their corresponding *t*-values) allowed us to

relate the dynamics of action representations to trial-to-trial variability in behavior (see *Multilevel Modeling* section for details). In all RSAs, we logit-transformed classification probabilities and further included a subject-specific "conjunction RT" model (i.e., a vector of z-scored, RTs, averaged for each subject and action constellation) as a nuisance predictor to reduce potential biases in decoding due to idiosyncratic differences in RTs among action constellations. We excluded resulting *t*-values that exceeded 5 SDs from means for each sample point, which excluded 0.12% and 1.32% of the entire samples Experiments 1 and 2 respectively. Resulting *t*-values were averaged within in 12 ms non-overlapping time samples.



Supplementary Fig. 1. Mean RTs of individual subjects for all action constellations in Experiment 1. Subjects-specific RT vectors were included as a nuisance predictor during RSA fitting.



Supplementary Fig. 2. Mean RTs of individual subjects for all action constellations in Experiment 2. Participants responded slowly for trials with the non-symmetric translation (clockwise and counterclockwise rules) compared to the symmetric rules (vertical and horizontal rules) as reported previously[7]. Subjects-specific RT vectors were included as a nuisance predictor during RSA fitting.

In Experiment 1, we constructed RSA models for the rules, stimuli, responses, and conjunctions (Fig. 1c). In Experiment 2, the conjunction model was separated for the rule-specific S-R conjunction model (rule-S-R conjunction) and the rule-independent S-R conjunction model (S-R conjunction; Fig. 5b). Complete orthogonalization of features could be established within each of two equal-sized subspaces of the entire space of action constellations, but not across the entire space. Therefore, we performed the RSA within each of these subspaces independently and subsequently averaged the results. Specifically, one subspace (G1 in Fig. 5) contained constellations with stimuli at the top-left or bottom-right corner (leading to a bottom-left or bottom-right response for all rules), whereas the second subspace (G2 in Fig. 5) contained trials with stimuli at the left-bottom or top-right corner (leading to a top-left or bottom-right response). Within each subspace, conjunctions were defined by the combination of four rules (vertical, horizontal, clockwise, and counterclockwise), two stimulus positions, and two responses, ensuring that each S-R link could occur in the context of two different action rules.

Non-Parametric Permutation Test

To test statistical significance of all time-resolved decoding results (Fig. 3a and Fig. 7a and Supplementary Fig. 3 and Supplementary Fig. 4) while accounting for multiple comparisons, we carried out nonparametric permutation tests using the single-threshold method[43]. For each feature, we computed permutation distributions of the maximum statistic for every sample point from -200 ms prior to the onset of the cue to 600 ms after the onset of the stimulus. Specifically, we first obtained classification results (and performed RSA for Fig. 3a and Fig. 7a) by decoding of data with randomly shuffled condition labels. We then performed a series of *t*-tests for every sample against the null level (i.e., the chance level). For the RSA results, the null level was 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 time points by comparing scores from the correct labels to the critical threshold, which was defined as the 99th (i.e., alpha =.01) of the largest member of maximum statistics in the permutation distribution of the corresponding variable.

Multilevel Modeling

We used multilevel linear modeling to analyze trial-by-trial variability in decoded representations and their relationship to RTs. The models estimated fixed effects of predictors as well as subject specific intercepts and slopes as random effects. For all statistical tests, the dependent variable (e.g., RSA scores or RTs) was prewhitened by the linear and quadratic trends of experimental trials and blocks. RTs were further log-transformed before the fitting. We performed statistical tests for a-priori selected time

intervals: cue-to-stimulus period from the onset of the cue to the onset of the stimulus (-300 to 0 ms for Experiment 1 and -500 to 0 ms for Experiment 2), early post-stimulus period (0 to 300 ms of the post-stimulus segment for both studies), and late post-stimulus period (300 to 600 ms of the post-stimulus segment for both studies). We predicted trial-to-trial RTs/RSA scores in the current trials with EEG signals from pre-stimulus and early post-stimulus periods in hopes of capturing processing prior to response execution (see also Supplementary Fig. 7 for results using signals aligned to the response onsets). The late post-stimulus interval was used to assess how partial-overlap costs are modulated by the strength of action representations developed during selection in *n*-1 trials. In addition, we separately performed a series of regressions to visualize changes in RT predictability—"impact" of moment-to-moment strength of decoded feature—by fitting models at each sample point without random slopes (Fig. 3b and Fig. 7b).

Supplementary Results

Behavioral Effects

Supplementary Table 1 and Supplementary Table 2 contain the results of Anovas of the RT and error effects as a function of the rule change/repeat factor and the response repeat/change factor in Experiment 1, and of the rule change/repeat factor and the S-R repeat/change factor in Experiment 2.

Supplementary Table 1.
Anovas of RTs/errors with the factors of rule or S-R repeat/change for Exp. 1

		RT			Error	
	F(1,19)	*P*	η^2	F(1,19)	*P*	η^2
Rule repeat/change (A)	3.34	.083	.149	.01	.933	<.001
Resp. repeat/change (B)	5.54	.029	.225	2.72	.116	.125
(A) x (B)	67.07	<.001	.779	21.41	<.001	.530

Supplementary Table 2.
Anovas of RTs/errors with the factors of rule or S-R repeat/change for Exp. 2

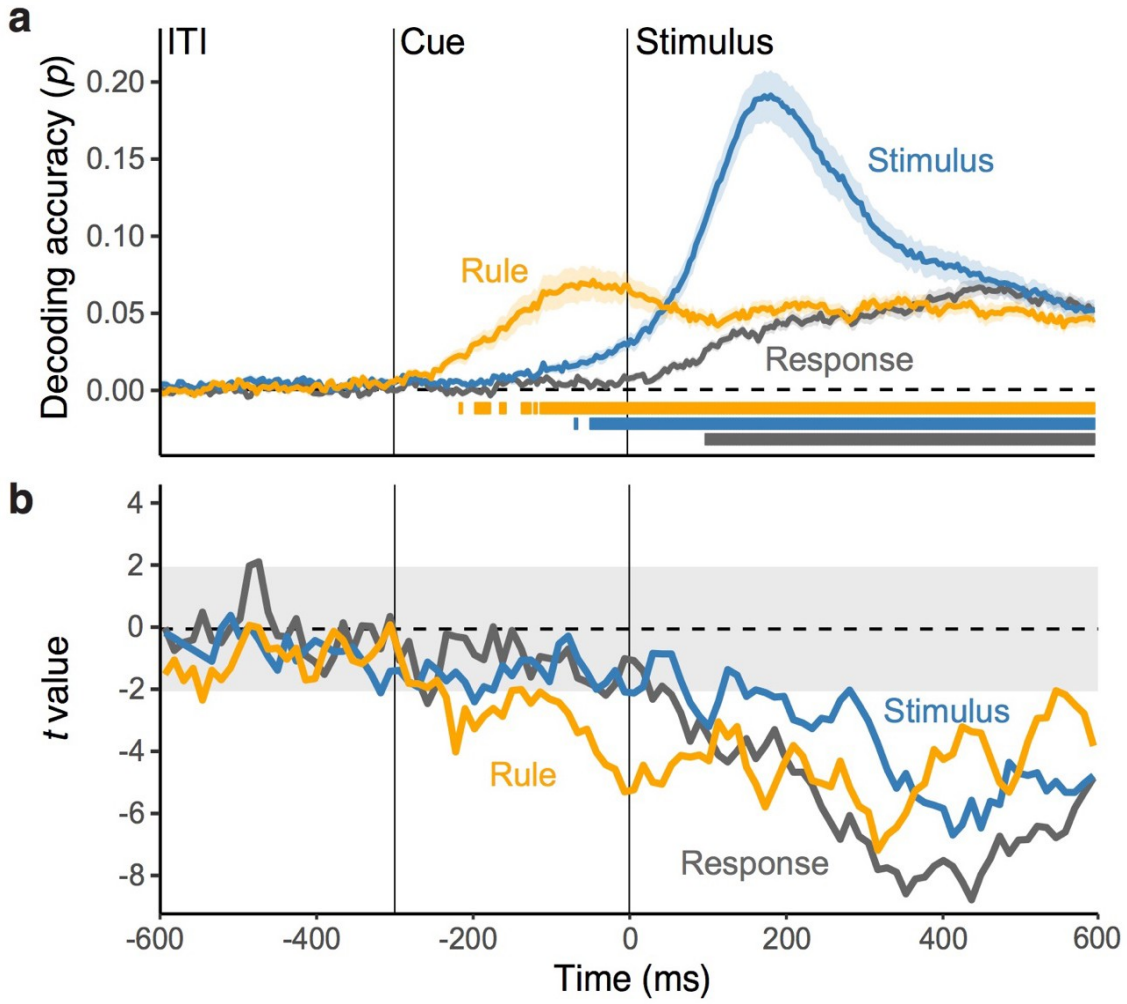
		RT			Error	
	F(1,20)	*P*	η^2	F(1,19)	*P*	η^2
Rule repeat/change (A)	5.08	.036	.202	.07	.791	.003
S-R repeat/change (B)	42.28	<.001	.679	1.72	.204	.079
(A) x (B)	25.56	<.001	.561	6.91	.002	.257

Decoding of basic, constituent features without RSA

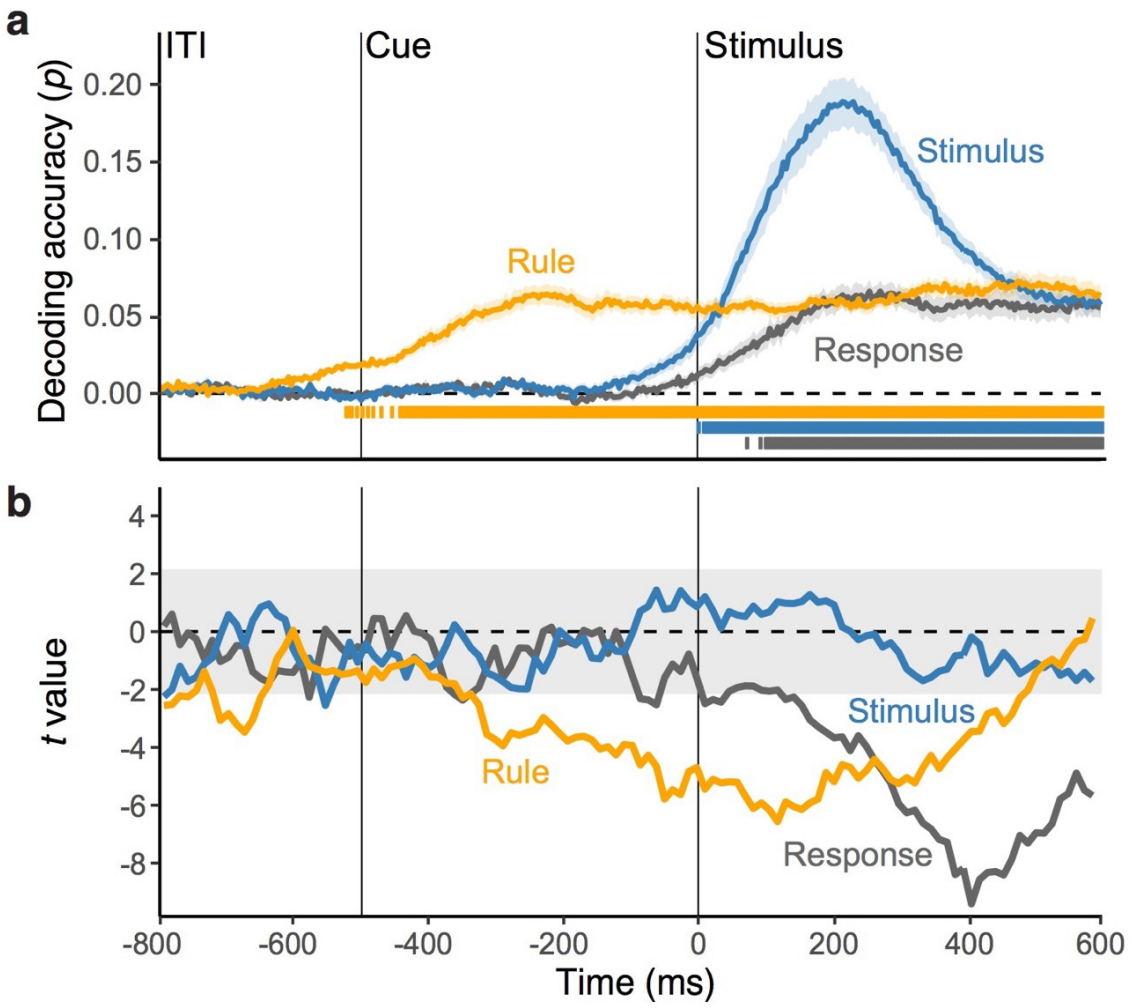
In the main paper, we used RSA analyses to distinguish conjunction representations from the representation of constituent features. In order to compare these results with a standard decoding approach, we also performed standard multivariate decoding analyses for each constituent feature independently. The analysis procedure (i.e., cross-validation, non-parametric permutation test, and subsequent multilevel modeling predicting trial-to-trial RTs) was identical to the method for RSA except for the following points: 1) final outputs of decoding analysis were classification probabilities (then logit-transformed) rather than RSA fit score, and 2) individuals-specific mean RTs of all action constellations were included as a control predictor in multilevel models of RTs.

Supplementary Fig. 3 and Supplementary Fig. 4 show the trajectories of classification probabilities of constituent features (rules, stimuli and responses) and their impact on trial-to-trial variability in RTs (see the inserts of Fig. 3a and Fig. 6a for the corresponding RSA results). The results were overall consistent with RSA results, when excluding the conjunction models (i.e., see inserts for Fig. 2a and Fig. 4a). The rule was activated during the pre-stimulus phase, followed by a strong expression of the stimulus, and finally by the response. This pattern directly replicates results using a more standard task-switching paradigm. These results also confirmed that our RSA approach produced

qualitatively similar results to the standard time-resolved decoding analysis for constituent features (when the conjunction model was excluded).



Supplementary Fig. 3. **a**, Average decoding accuracy of constituent features over time, derived from a standard decoding analysis in Experiment 1. Shaded regions specify the standard error around the mean. Squares below lines denote the significant time points correcting for multiple comparison using a non-parametric permutation test. **b**, Time-course of *t* values from multilevel, linear models predicting the variability in trial-to-trial RTs, using single-trial classification probability of each feature as predictors simultaneously.



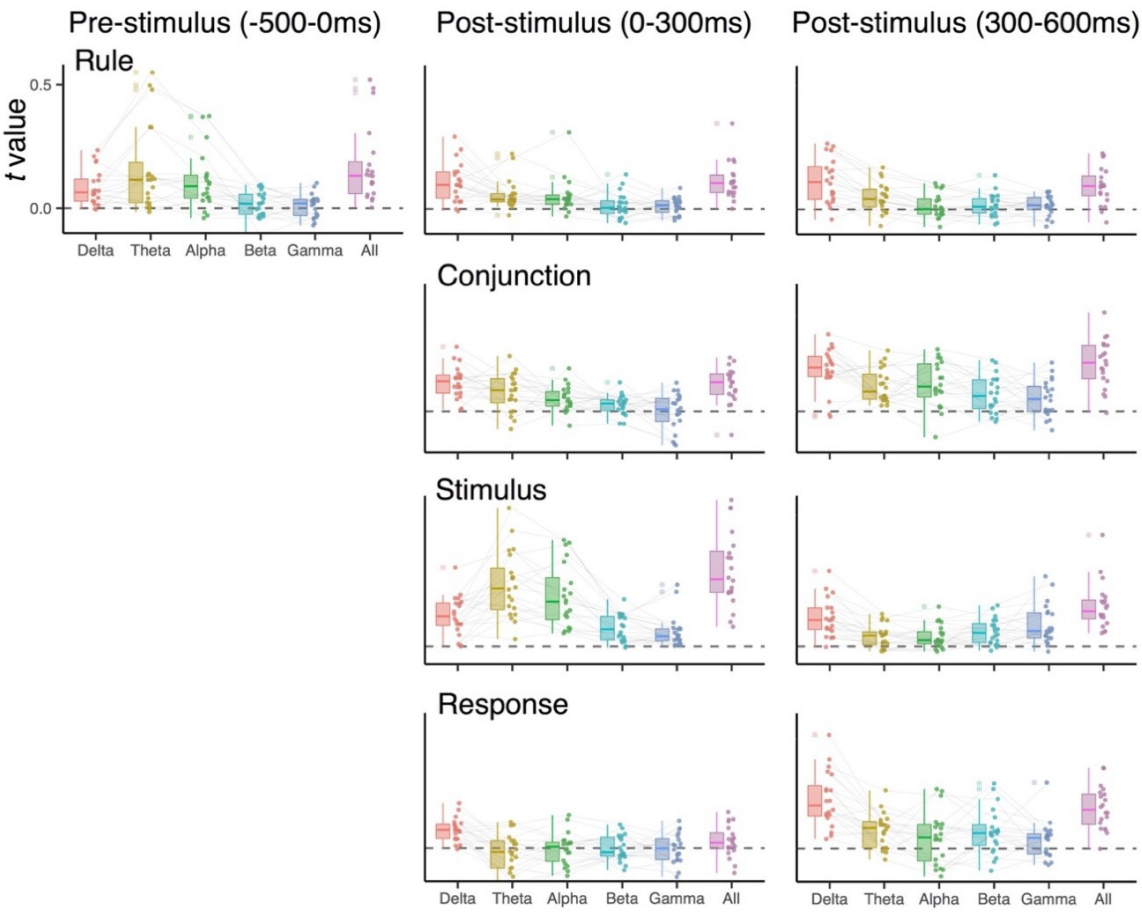
Supplementary Fig. 4. **a**, Average decoding accuracy of constituent features over time, derived from a standard decoding analysis in Experiment 2. Shaded regions specify the standard error around the mean. Squares below lines denote the significant time points correcting for multiple comparison using a non-parametric permutation test. **b**, Time-course of *t* values from multilevel, linear models predicting the variability in trial-to-trial RTs, using single-trial classification probability of each feature as predictors simultaneously.

RSA using frequency-specific EEG activity

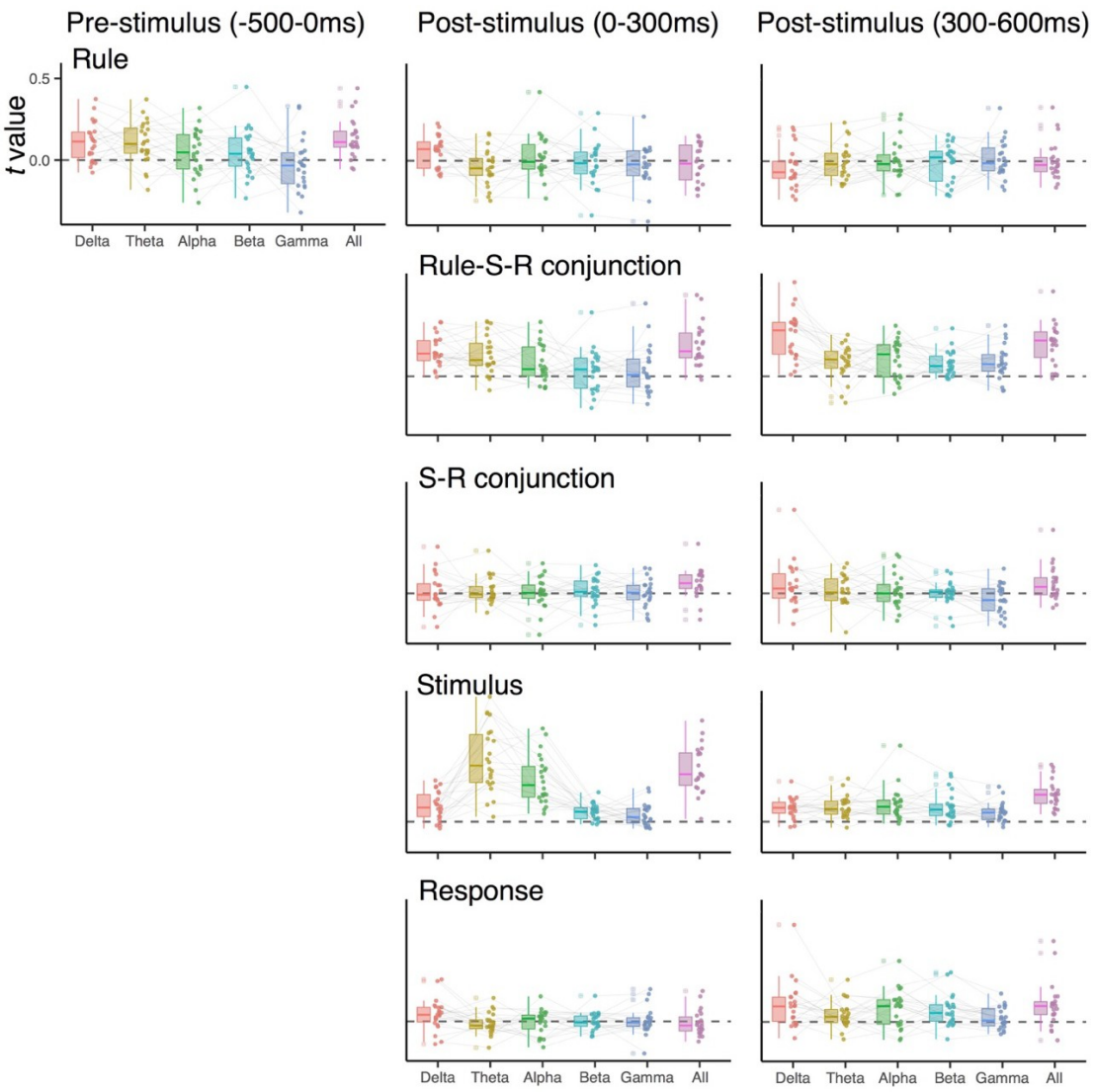
Previous studies showed that specific control representations are encoded in the frequency-specific, rhythmic EEG activity (e.g., the ordinal position codes in the theta-band (4-7 *Hz*)[45]. As an exploratory analysis, we analyzed how different frequency-bands contribute to the decoding of both conjunctions and constituent features. For this, we replicated the combination of decoding and RSA analyses separately for the delta (1-3

Hz), theta (4-7 *Hz*), alpha (8-12 *Hz*), beta (13-30 *Hz*) and gamma (31-35 *Hz*) frequency bands. To reduce the influence of temporal smearing, which could differ across frequency-bands[38], we averaged data over a-priori selected time intervals (i.e., pre-stimulus, early and late post-stimulus phase) prior to the training of decoders. Other steps in the analysis were identical to the one described in *Representational Similarity Analysis* section.

Supplementary Fig. 5 and Supplementary Fig. 6 summarize RSA scores of different action features among individuals for each time interval, compared to the results using all five frequency-bands. Overall, participants exhibited considerable variability in terms of the frequency-bands in which specific information was expressed. However, the following points seem to be consistent across individuals: 1) stimulus positions in early post-stimulus interval were coded in the theta and the alpha-band, a finding that is consistent with earlier decoding results [14], and 2) conjunctions (i.e., rule-S-R conjunctions in Experiment 2) tended to be expressed most strongly in the delta band. The fact that the neighboring theta-band activity only weakly contributed to the conjunctions emphasizes the special role of delta-band activity. We had no a-priori predictions about frequency-specific effects. However, we note that there are recent reports suggesting that delta-band frequencies carry information about abstract decision variables [47]. Using a similar "search-light" method across electrodes (instead of frequency-space), we also checked for relative contribution of local scalp regions for expression of representations that are more localize consistent across individuals. However, we found the patterns with which representations were expressed to be highly idiosyncratic (see Kikumoto and Mayr [45] for similar results in hierarchical serial-control task).



Supplementary Fig. 5. Experiment 1: RSA scores when decoding EEG signals in specific frequency ranges (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, and 1-35 *Hz* for all). EEG signals were averaged over pre-stimulus (-300 to 0 ms), early post-stimulus (0 to 300 ms) and late post-stimulus (300 to 600 ms) time intervals before the decoding analysis.

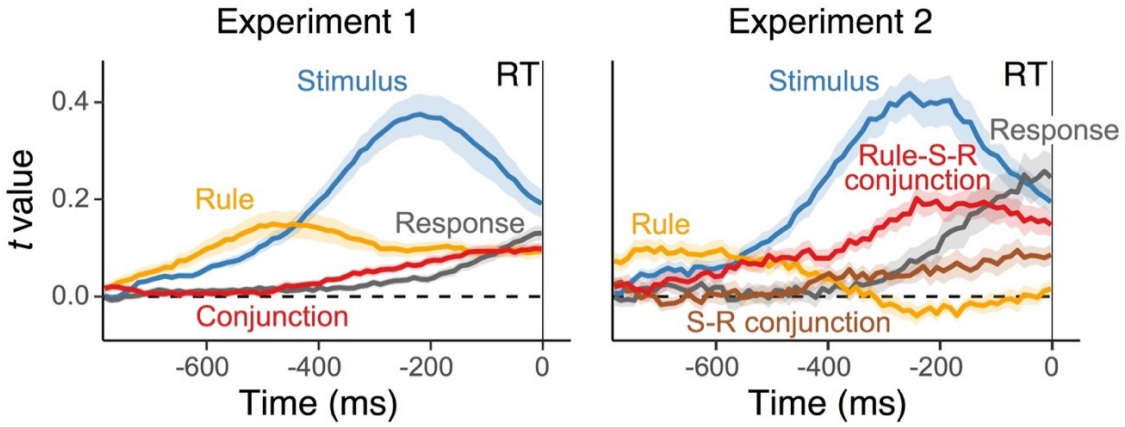


Supplementary Fig. 6. Experiment 2: RSA scores when decoding EEG signals in specific frequency ranges (1-3 *Hz* for the delta-band, 4-7 *Hz* for the theta-band, 8-12 *Hz* for the alpha-band, 13-30 *Hz* for the beta-band, 31-35 *Hz* for the gamma-band, and 1-35 *Hz* for all). EEG signals were averaged over pre-stimulus (-300 to 0 ms), early post-stimulus (0 to 300 ms) and late post-stimulus (300 to 600 ms) time intervals before the decoding analysis.

RSA time-aligned to responses

The trajectory of RSA fit scores, summarized in Fig. 3a and Fig. 6a, revealed how action-relevant representations unfold in reference to the onset of a stimulus. However, these decoding results contain a limited number of samples which responses were executed during the post-stimulus interval (average RT: M = 490 ms, SD = 1.46 ms in

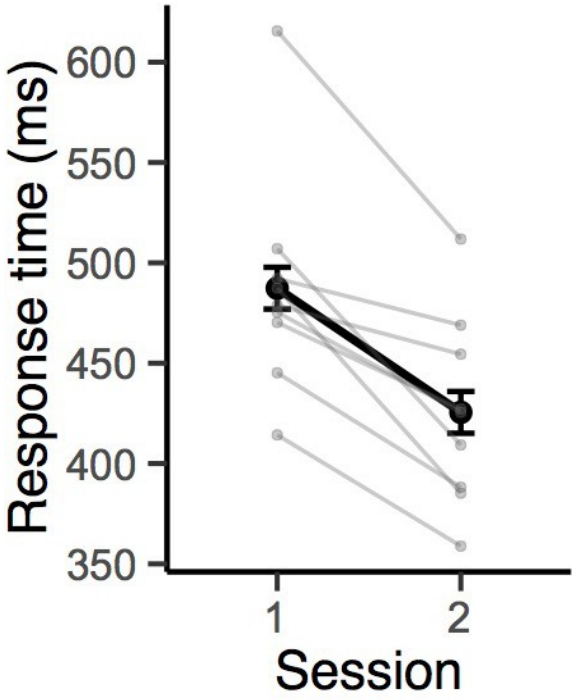
Experiment 1; *M* = 490 ms, *SD* = 1.67 ms in Experiment 2). To ensure that our stimulus-locked decoding results are not affected by such early responses, and to provide information about how representations unfold relative to responses, we performed an additional RSA using EEG data that are aligned to the onset of each trial's response. Supplementary Fig. 7 shows the corresponding time-course of RSA scores for Experiments 1 and 2. These results are highly consistent with stimulus-locked results in showing early peaks for rule, followed by stimulus, then conjunction, and finally response representations, which peaked just before responses were executed.



Supplementary Fig. 7. Average, single-trial RSA scores for each representation, using EEG signals aligned to the onset of trial-to-trial response events. Shaded regions specify standard error of the mean.

Cross-session RSA

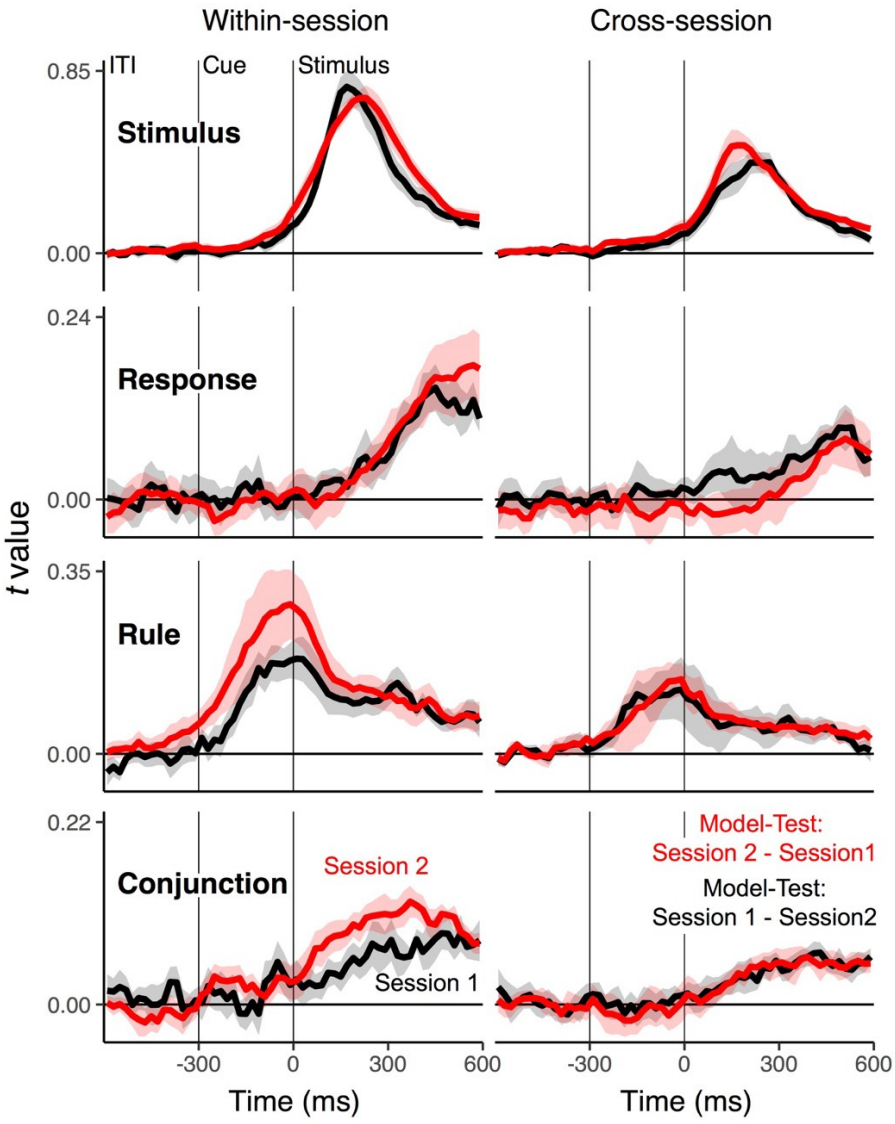
Currently, we know relatively little about within-subject consistency of EEG-decoded, complex action representations across longer temporal intervals. In particular, given that the neural patterns that we decoded were highly idiosyncratic (see Section ***RSA using frequency-specific EEG activity***), it is important to explicitly test the stability of such patterns within individuals. We were able to recruit 9 out of the 20 participants from Experiment 1 for a second session, 6 months after the first session.



Supplementary Fig. 8. Individuals' mean RTs and group average RTs (black) between the first session and the second session at least 6 months later (n = 9) for Experiment 1. Error bars specify 95% within-subject confidence intervals.

Participants responded significantly faster in the second session than in the first session, RTs: $F(1,8) = 29.05$, $MSE = 735.54$, $p < .001$, $\eta p^2 = .78$; errors: $F(1,8) = .40$, $MSE < .001$, $p = .54$, $\eta p^2 = .05$ (Supplementary Fig. 8). In order to perform cross-session RSAs we trained decoders using data within each session separately. Supplementary Fig. 9 (left column) summarizes differences in RSA scores of each representation across sessions. Both the pre-stimulus rule representation and the conjunction in the stimulus-to-response phase were enhanced in the later session. Then, decoders that were trained separately within sessions were applied to EEG data of the other session at matching time points. As shown in Supplementary Fig. 9 (right column), all features, including the conjunctive representation, showed robust generalizability across sessions. This pattern indicates that the neural representation of action-relevant representations is remarkably stable. An important implication of this result is that in future research such techniques

can be applied to analyze how action-relevant representations are shaped through experience or consolidation[48].



Supplementary Fig. 9. Average, single-trial RSA fit scores for conjunctions and constituent features across sessions (*n* = 9). *Left panels*: Time-course of RSA scores training decoders. *Right panels*: RSA scores from cross-decoding across sessions. Decoders are trained with EEG data from the session 1 (shown in red) or session 2 (shown in black) and applied on data from the counterpart session to test their generalizability. Shaded regions specify 95% within-subject confidence intervals.

CHAPTER III.

ACTION STOPPING

Introduction

Even for a simple goal-directed action, such as kicking a soccer ball to a teammate, various sensorimotor features—the location of the ball, the presence of opponent players and teammates, as well as the abstract action rules (e.g., "kick softly when the grass is wet")—need to be adequately represented. Prominent theories of action control suggest that rather than being handled independently, action-relevant features are integrated within a common representational space during selection. Specifically, event-file theory[1,2] posits that an action becomes executable only once all task-relevant features are integrated within conjunctive representations (i.e., event files).

If conjunctive representations are necessary, and maybe even sufficient precursors of successful action, it follows that the pathway to controlling a given action needs to lead through these representations. For example, when an opponent defender suddenly blocks the targeted teammate, the kicking action needs to be canceled. Action inhibition, including its neural underpinnings, has been well characterized using variants of the stop-signal paradigm[3–9]. Yet, it is currently an open question how the stopping affects different representations for the planned action—which of these action codes are targeted by the stopping process. In theory, the stopping process might occur simply by suppressing the response representations that directly map onto at various levels of motor control pathways[10–13]. However, assuming conjunctive representations are indeed critical for action control, the cancellation of an initiated action should require the suppression of the corresponding conjunction—the integrated, holistic representation of the action plan binding even abstract cognitive features.

Kikumoto and Mayr (2019) recently applied a time-resolved representational similarity analysis (RSA)[14] to the EEG signal in order to track the presence of conjunctive and constituent feature representations during action selection, on the level of individual trials. 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 predictor of trial-to-trial variability in RTs even compared to the constituent (but task-relevant) rule, stimulus, or response representations—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)[15,16] with an occasional stop signal[3,17,18]. As in Kikumoto and Mayr (2019), we used RSA to track action-relevant representations. In Experiment 1, the stop signal was presented 100 ms after the stimulus onset, which is early enough for successful stopping on most trials. This allowed us to determine which action representations are suppressed in a reactive manner on stop-trials, relative to go-trials. In Experiment 2, stop-signal timing was adjusted via staircase-tracking procedure[4] based on participants' trial-to-trial stopping accuracy[18]. This allowed us to replicate and extend the results of Experiment 1 by relating the strength of decoded representations to the accuracy of stopping on a trial-by-trial basis. Across both experiments, we found strong evidence that stopping selectively suppresses conjunctive representations. In Experiment 2, we also found that the strength of conjunctive representations at the time the stop-signal arrives, is a unique predictor of stopping success.

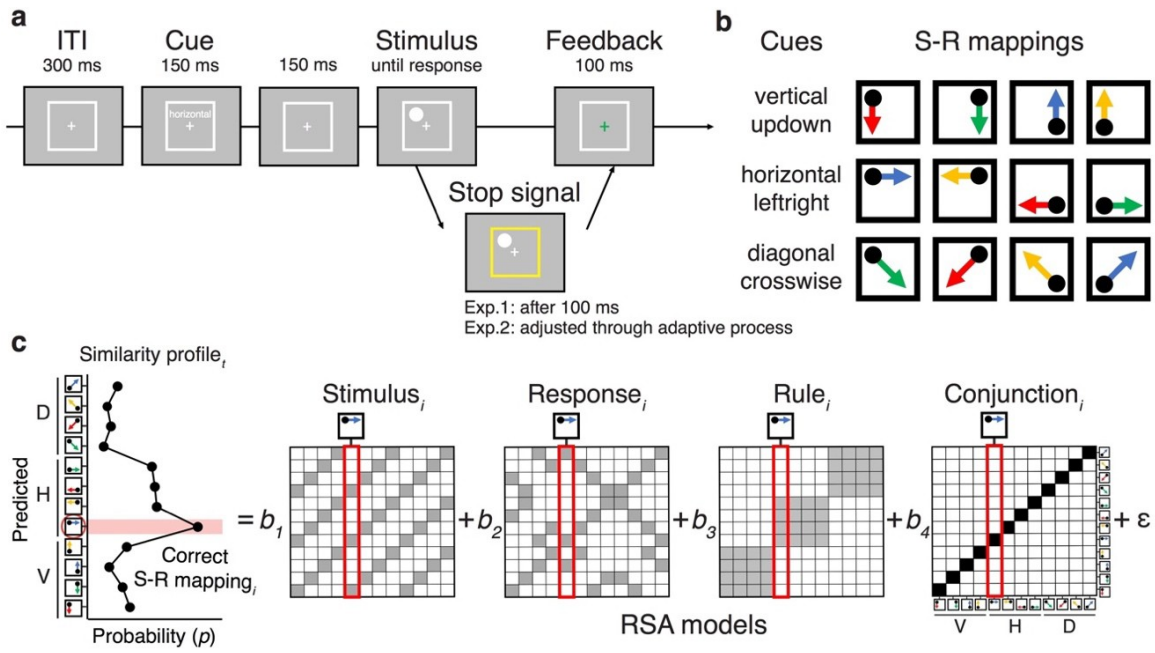


Fig. 1. a, Sequence of trial events in the rule-selection task combined with the stop-signal task for both Experiment 1 and 2. **b**, Spatial translation rules mapping specific stimuli to responses. Two different cue words were used for each rule. **c**, Schematic steps of the representational similarity 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.

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 (Exp.1 in Kikumoto & Mayr, 2019). 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 the fixed starting point.

Table 1.
Average behavioral performance in go and stop-trials

	RT (Go)	Error (Go)	p(respond|signal)	Failed Stop RT	Error (Stop)	SSD	SSRT
Exp.1	504 (62.5)	3.24 (1.77)	11.6 (8.85)	405 (32.2)	2.36 (2.66)		
Exp.2	676 (139)	3.55 (3.64)	44.1 (4.84)	552 (91)	2.83 (2.28)	327 (103)	272 (54.8)

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[15,19]. Importantly, conjunctive information was present throughout the entire response-selection period[15]. We also replicated our findings that trial-to-trial variability in the conjunctive representations robustly predicted go-trial RTs (Supplementary Table 1), over and above other representations of constituent features, indicating the critical role of integrated representations for the execution of goal-directed actions.

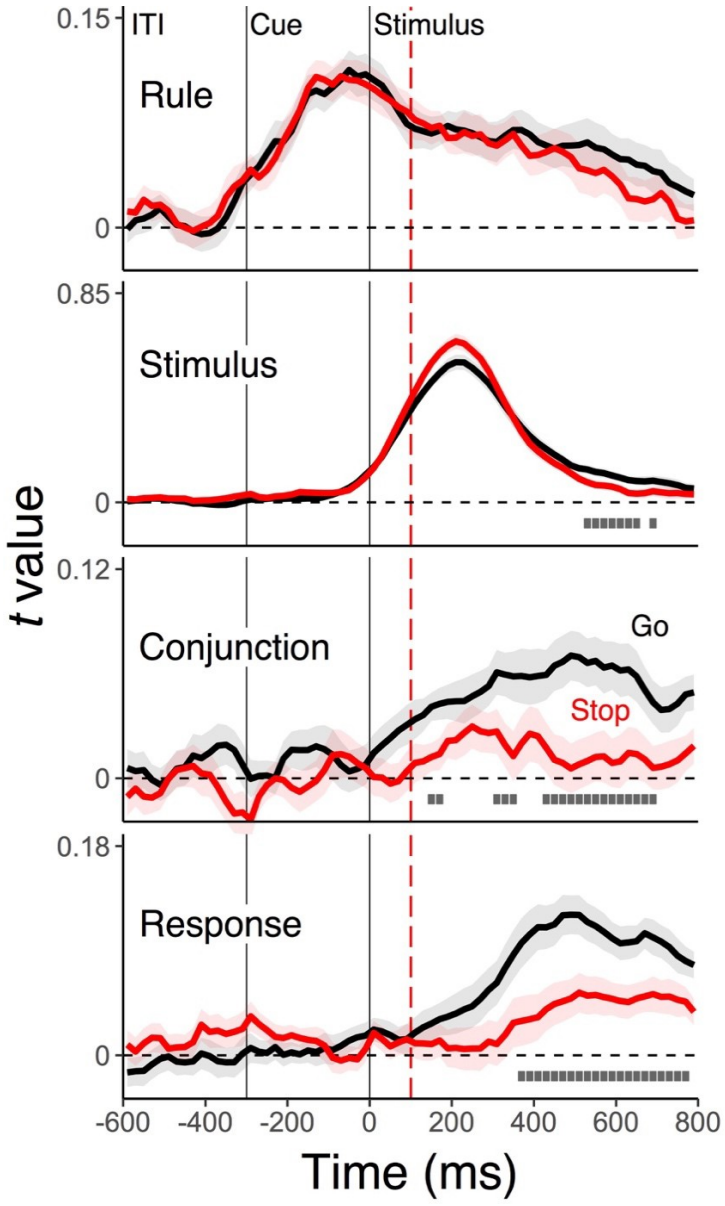


Fig 2. Average, single-trial *t*-values derived from the RSA (see Fig. 1c) for each of the basic features (rule, stimulus, and response) and the conjunction, separately for go- (black) and stop-trials (red). Shaded regions specify the 95% within-subject confidence intervals. The vertical, red dashed line marks the onset of the stop-signal at 100 ms after the stimulus onset.

Our main goal in Experiment 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 the suppression effect on conjunction 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.

When we do not explicitly model the conjunctive representations, the suppression effect on the basic features was substantially conflated (Supplementary Fig. 3), highlighting the importance of including the conjunction model to understand action control (see also Kikumoto & Mayr, 2019). Thus, overall, these results are consistent with the prediction of event-file theory: conjunctive representations drive action selection and serve as the key target of the stopping process of planned actions.

Experiment 2

The results of Experiment 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 and stopping behavior. In Experiment 2, we therefore attempted to link trial-to-trial stopping success to the variability in action representations by using 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 ***Stimuli, Tasks and Procedure*** and ***Stop-signal reaction time (SSRT)*** section 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)[18]. 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).

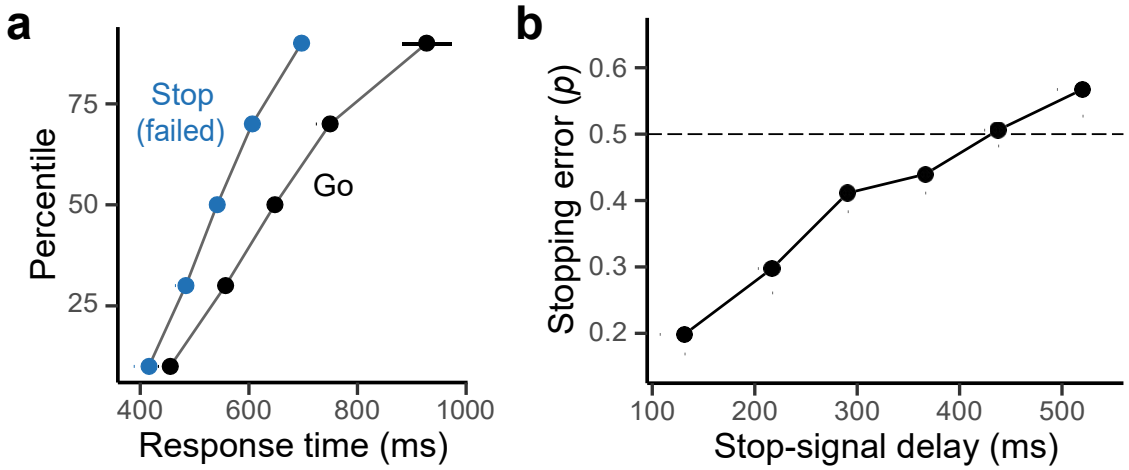


Fig. 3. **a**, Vincentized mean response times (RTs) for go-trials and failed stop-trials. **b**, Mean stopping failures as a function of changes in stop-signal delays (i.e., inhibition functions). Error bars specify 95% within-subject confidence intervals.

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), differentiating between go-trials, failed stop-trials and successful stop-trials. Replicating the results in Experiment 1, the strength of conjunctive representations was the most robust predictor of trial-to-trial RTs in go-trials (Supplementary Table 1). We again found that rule and stimulus representations showed little effects of stopping (excluding the conjunction model also yielded similar results to Exp.1; Supplementary Fig. 4). 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 a typical, post-stop-signal effects was detected only on the response representation 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 when the stop process is initiated determines stopping success.

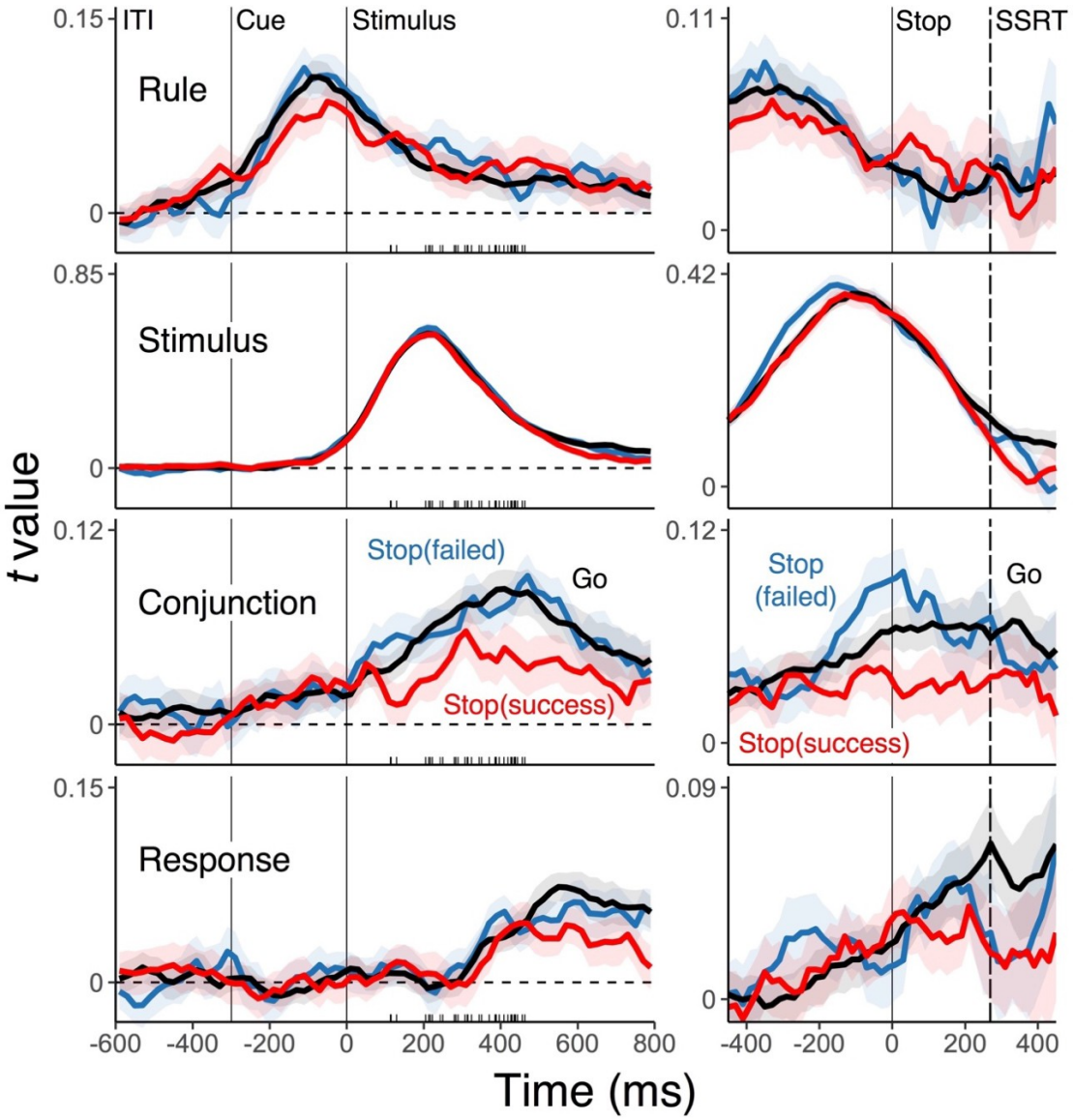


Fig 4. Average, single-trial *t*-values associated with each of the basic features (rule, stimulus, and response) and their conjunction derived from the RSA, separately for go- (black), successful stop-trials (red), and failed stop-trials (blue). The left panels show the results that are aligned to the stimulus onset and the right panels correspond to the results that are aligned to the stop-signal onset. Shaded regions specify the 95% within-subject confidence intervals. Tick marks on the x-axis mark individuals' average stop-signal delay.

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 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 due to 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 other features, strongly predict single-trial stopping failures prior to the onset of the stop-signal (Fig. 5). These results were robust when we account for 1) the trial-to-trial SSDs (Fig. 5) and 2) the premature responses that occurred before the stop-signal onset (Table 2). 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 3). Together, these results indicate that the conjunctive representations are a critical determinant of successful stopping of actions: the integrated representations of action features strongly drive actions, which in turn makes the stopping of actions more challenging.

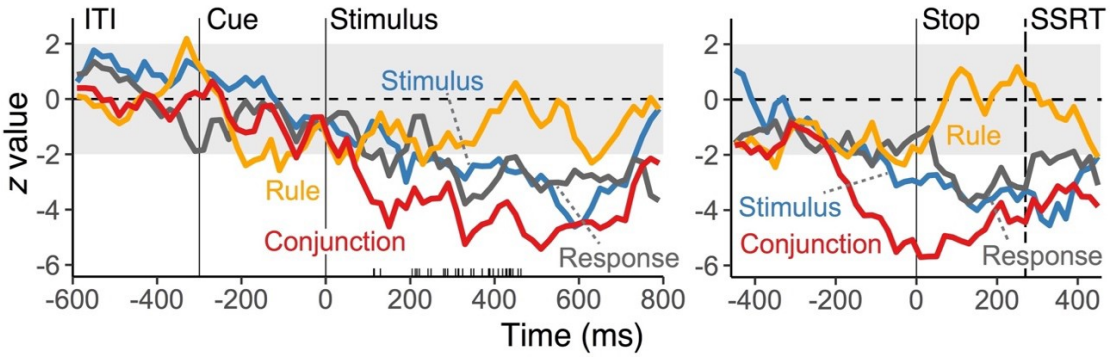


Fig. 5. Time-course of *z* values from multilevel, logistic regression models predicting the variability in trial-to-trial stopping failures in the stop-trials (the "impact" of representations on stopping behavior), using RSA scores of all features and trial-to-trial SSDs as the simultaneous predictors.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

Table 2.
Predicting trial-by-trial stopping accuracy using the strength of decoded representations.

		Pre-Stop-Signal		Post-Stop-Signal	
Model	Decoded 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

Table 3.
Predicting trial-by-trial stopping accuracy using the strength of decoded representations, using both pre-stop-signal and post-stop-signal intervals.

Time	Decoded variable	*b* (*se*)	*t* value
Pre-stop	Rule	-.084 (.036)	-2.31
	Stimulus	-.035 (.036)	-0.93
	Response	-.038 (.040)	-0.97
	Conjunction	-.126 (.049)	-2.60
Post-stop	Rule	.016 (.044)	0.36
	Stimulus	-.104 (.041)	-2.53
	Response	-.093 (.041)	-2.30
	Conjunction	-.194 (.049)	-3.93

Discussion

Goal-directed actions usually rely on multiple feature dimensions of the current task environment. Event-file theory proposes that these different action features come together within conjunctive representations to guide actions[1,20]. In our previous work[15], we had reported direct, 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 kicks in, 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, 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 the 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, Fig. 4, and Fig. 5). Other representations (e.g., rule and stimulus information) remained intact or showed very minor suppression only after the completion of the stopping process, as indicated by the SSRT (Fig. 4 and Supplementary Fig. 4). In particular, the representation of the rule remained unaffected by the stopping process. A potential functional benefit of selective suppression could be that by retaining the abstract rule information of the initiated action, actions can be easily re-implemented, once the reason for stopping the initial action has been removed.

Conjunctive representations that integrate disparate features, including abstract rule information, must be 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[21–25]. 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 the main target of inhibition has implications for the suppression-induced, long-term memory effects (e.g., Anderson and Green, 2001[26]). 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 the 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 effect of a reactive inhibition process (Table 2 and 3). This pattern strongly suggests that the strength of conjunctive representations is a necessary, 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 across trials. A strong emphasis on initiating action, accompanied by a strong conjunction, may cause the failures to trigger the stop process altogether[4,27,28]. 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 the conjunction[29,30]. 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[31,32]. In any case, our results clearly indicate the pre-stop-signal state of action representations must be taken into account to fully understand subsequent reactive inhibition and stopping.

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 main driver of successful action implementation.

Method

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 Experiment 1 and 36 out of 38 for Experiment 2. In Experiment 2, 3 participants were further excluded because of failures to stop in excess of 75% [18].

Stimuli, Tasks and Procedure

The task combined a variant of rule-based action task[16] with the stop-signal paradigm[3,17,18]. 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 Experiment 1, the stop-signal appeared 100 ms after the stimulus onset. In Experiment 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 consisted of a 16 seconds-interval 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 Experiment 1, and

1576 (*SD* = 162) and 773 (*SD* = 75) for Experiment 2. Throughout the experimental session, participants were reminded to make responses as accurate and fast as possible and not to "wait for" the stop-signal. In Experiment 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 incentive, 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; Pelli 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 (SSRT)

In Experiment 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 *n*th 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

EEG data was first segmented by 18.5 second intervals to include all trials within a block (see the Appendix A. EEG RECORDING AND PREPROCESSING). After time-frequency decomposition was performed (see the Appendix B. TIME-FREQUENCY ANALYSIS), 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.

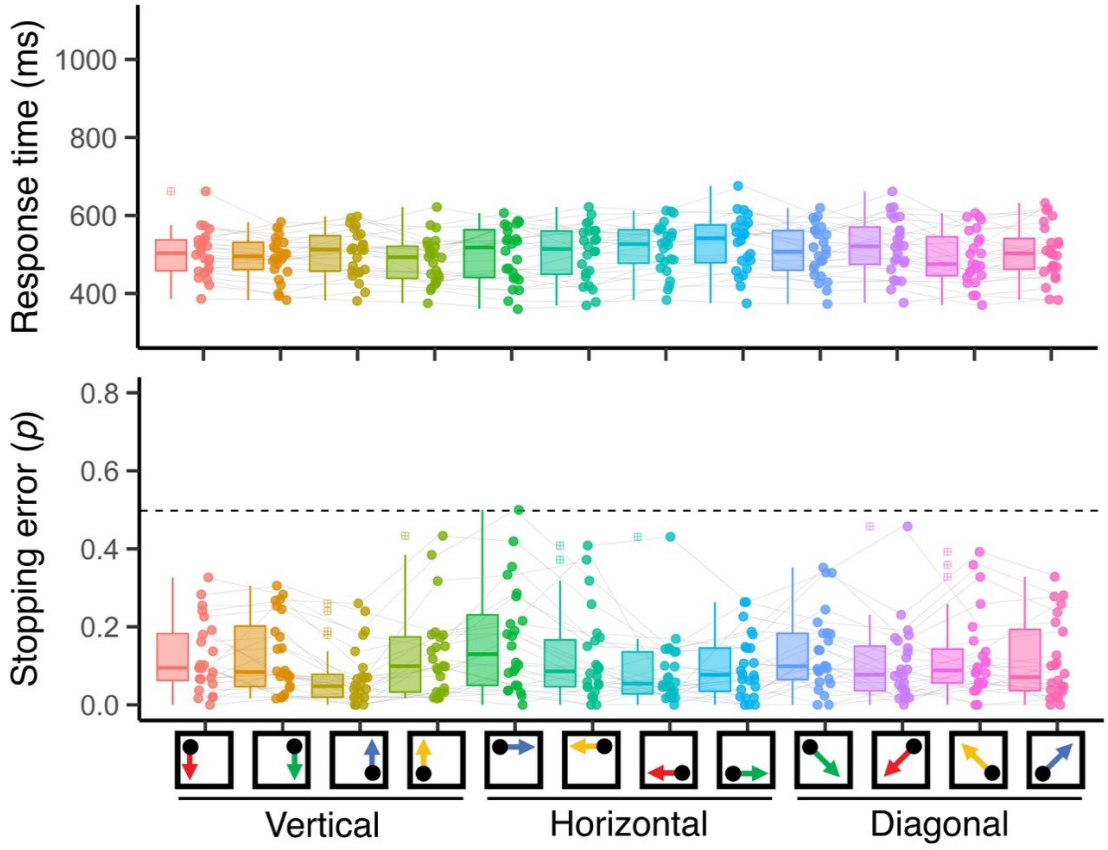
Representational Similarity Analysis

The decoding analysis in the current study follows Kikumoto and Mayr (2019): 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 different action constellations. Specifically, we performed a penalized linear discriminant analysis using the caret package in R [34]. 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*

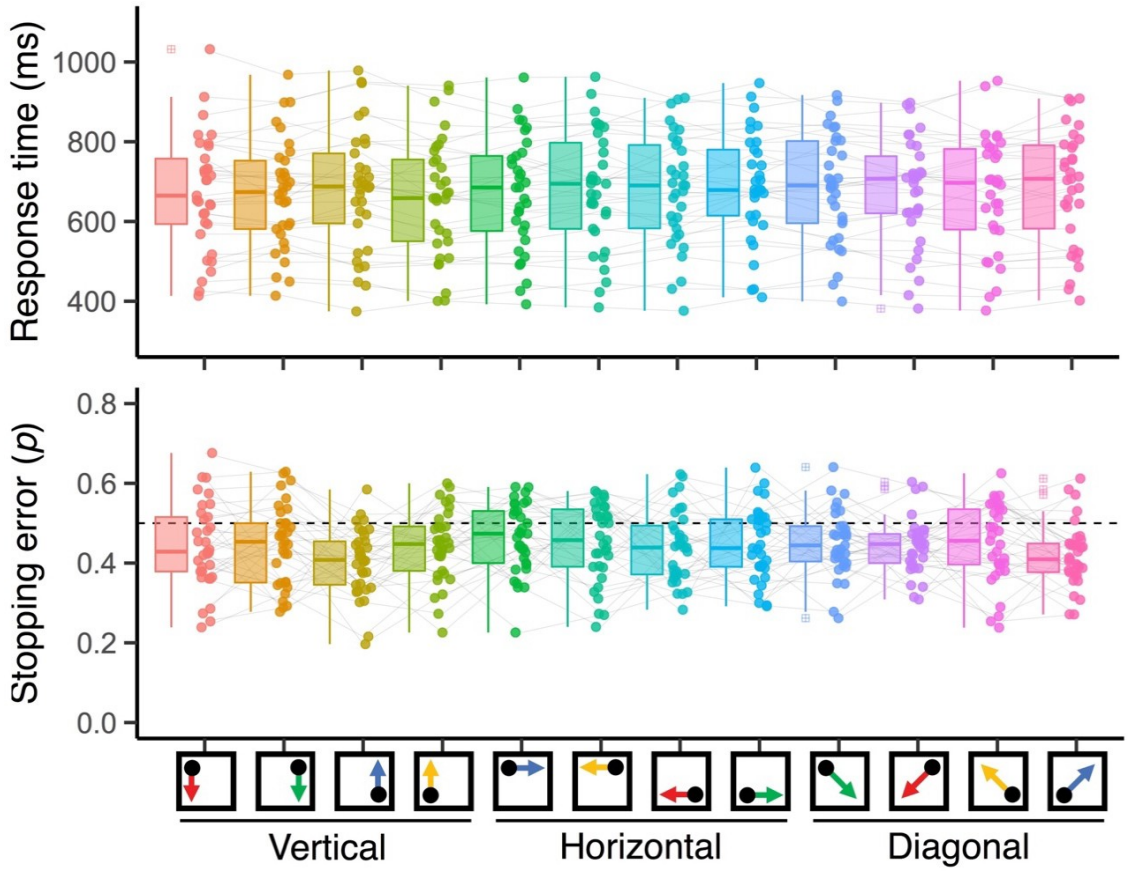
for the beta-band, 31-35 *Hz* for the gamma-band), generating 100 features (5 frequency-bands X 20 electrodes) to train decoders. Within individuals, these data points were z-transformed across electrodes at every sample to remove the effects that uniformly influenced all electrodes. We used a *k*-fold repeated cross-validation procedure to evaluate the decoding results[36], 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 (Supplementary Fig. 1 and 2).



Supplementary Fig. 1. Mean RTs and stopping error individual subjects for all action constellations in three different rules. To control potential differences covarying with average RTs/stopping errors, subjects-specific RTs/errors vectors were included as a nuisance predictor during RSA fitting.



Supplementary Fig. 2. Mean RTs and stopping error individual subjects for all action constellations in three different rules. To control potential differences covarying with average RTs/stopping errors, subjects-specific RTs/errors vectors were included as a nuisance predictor during RSA fitting.

In both experiments, decoders were trained with the stimulus-aligned EEG signal. In Experiment 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 Supplementary Fig. 4). 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 Experiment 1, we used nonparametric permutation tests with a single-threshold method [35] 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 Experiment 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 2 and 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 (Supplementary Table 1). Here, RTs were log-transformed and trials with response errors were excluded.

Supplementary Results

Predicting Single-trial Go Responses

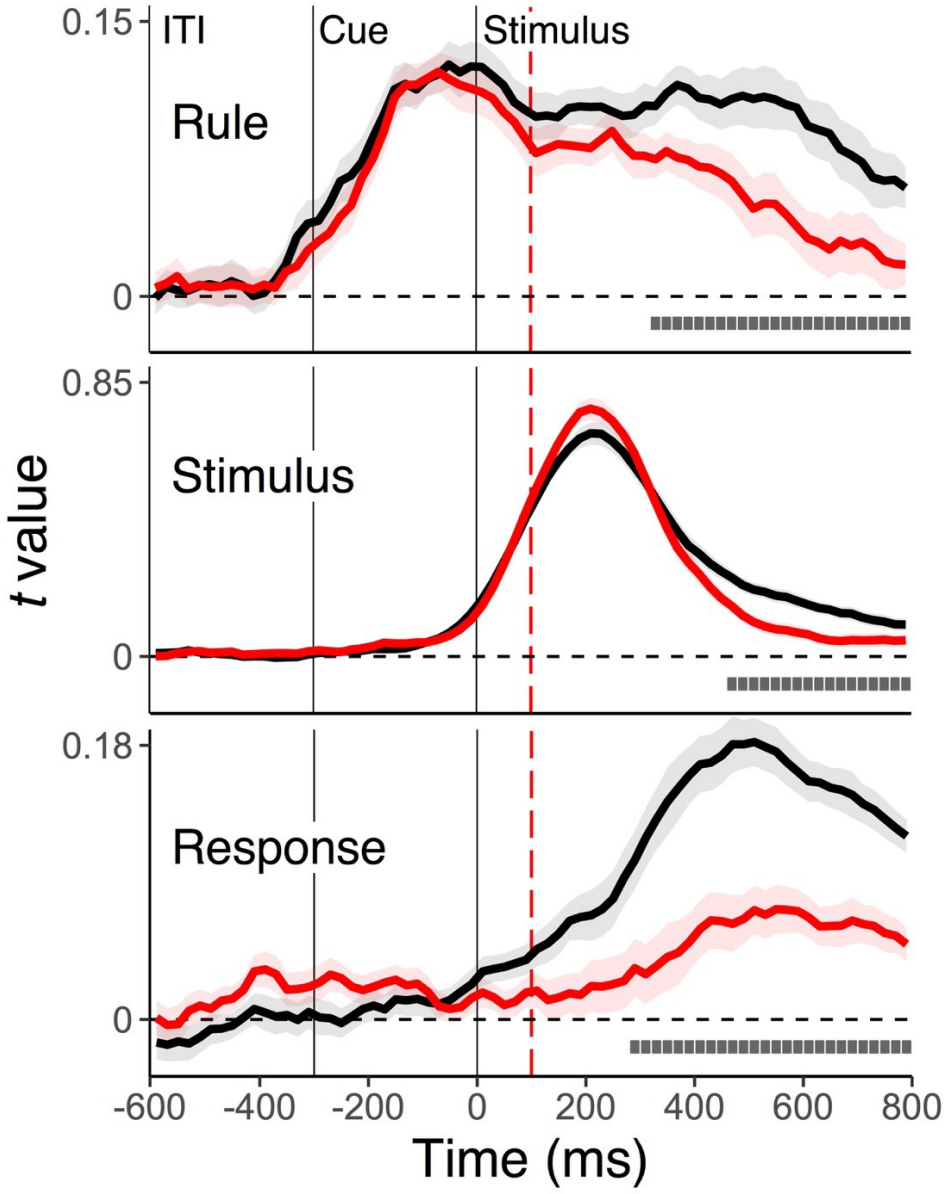
Supplementary Table 1 contain the results of multilevel modeling of single-trial RTs in go-trials, using all constituent feature and conjunctive representations as simultaneous predictors (see ***Multilevel modeling*** section for detail).

Supplementary Table 1.
Predicting trial-by-trial RTs in go-trials using the strength of decoded representations.

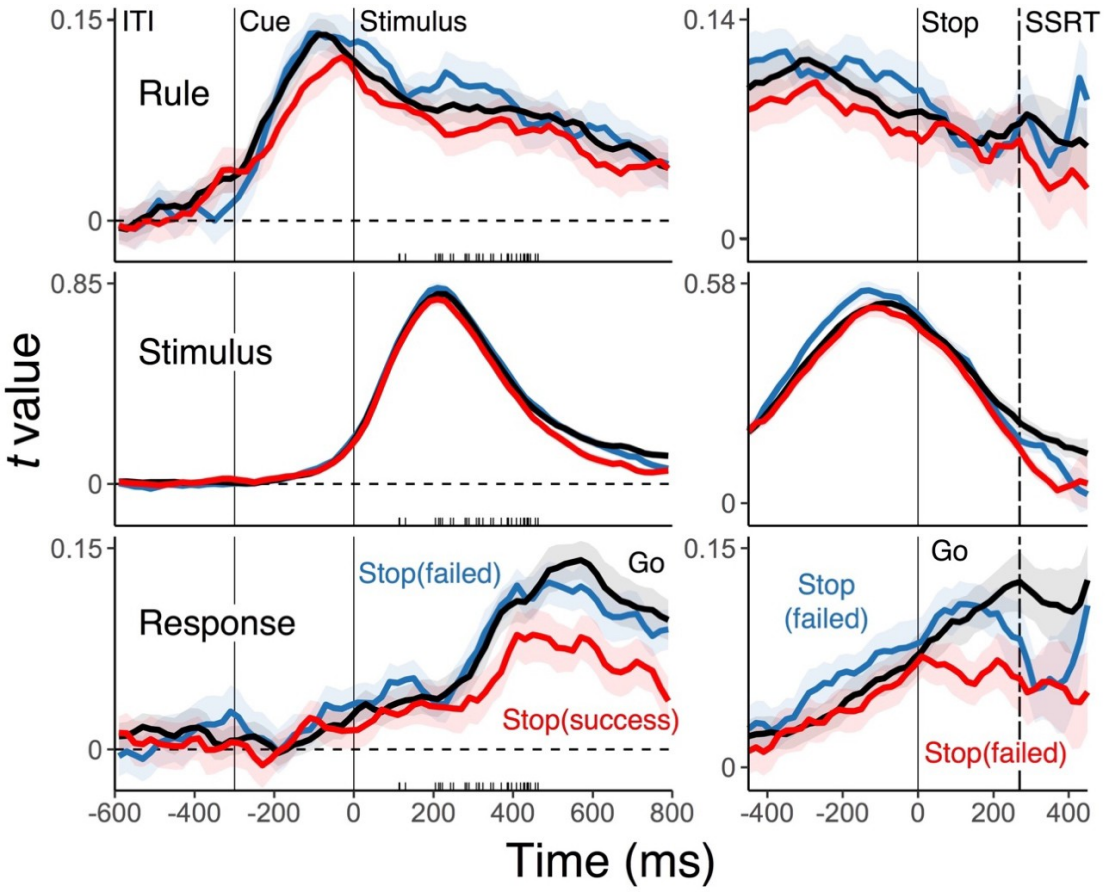
	Exp.1		Exp.2	
Decoded variable	*b* (*se*)	*t* value	*b* (*se*)	*t* value
Rule	-.025 (.012)	-2.15	-.013 (.006)	-2.08
Stimulus	-.015 (.009)	-1.59	-.038 (.009)	-4.12
Response	-.016 (.009)	-1.85	-.032 (.009)	-3.55
Conjunction	-.042 (.012)	-3.61	-.057 (.010)	-5.69

RSA without the conjunction model

Supplementary Fig. 3 and Fig. 4 show the RSA results excluding the conjunction model. The same decoding procedures and subsequent results (as shown in the Fig. 2 and Fig. 4) were used to perform RSA for comparison. In general, a pattern of results is similar to the earlier results (Chapter II. Fig.3 inset, Supplementary Fig. 3, and Supplementary Fig. 4). However, the suppression effect that is in fact driven by the conjunctive representations appear on the rule and stimulus information. Note if there is no conjunctive representations, addition of the conjunctive model should not change the results.



Supplementary Fig. 3. Average, single-trial *t*-values from RSA associated with each of the basic features (rule, stimulus, and response) without the conjunction model, separately for go-(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.



Supplementary Fig. 4. Average, single-trial *t*-values from RSA associated with each of the basic features (rule, stimulus, and response) without the conjunction model, separately for go-(black), successful stop-trials (red), and failed stop-trials (blue). The left panels show the results that are aligned to the stimulus onset and the right panels correspond to the results that are aligned to the stop-signal onset. Shaded regions specify the 95% within-subject confidence intervals. Tick marks on the x-axis mark individuals' average stop-signal delay.

CHAPTER IV.

FUTURE DIRECTIONS

Across four experiments, we characterized the unique functional role of nonlinear, conjunctive representations during action control. Our results are generally consistent with an "integration account" of action selection[1–5]. In principle, conjunctive representations can guide a variety of goal-directed behavior by combining more basic (i.e., generalizable) task representations. Integration of action features allows our neurocognitive system to be efficient and flexible at the same time, because abstract representations that are detached from the specific action contexts can be configured on ad hoc basis for different goals. Thus, the formation of integrated representations could be one of the canonical functions in encoding of task representations in humans. I had discussed issues that were specific to the individual chapters in the corresponding *Discussion* sections. In the remainder of this chapter, I present three avenues for future research, where EEG-based decoding techniques hold promise for significant theoretical progress.

Active maintenance and population coding

In many real-world situations we need to plan an action and maintain it in an active state until we can actually execute it. Often this also requires the need to exclude potentially interfering information that is not relevant for the current action plan. Evidence suggests that task sets for upcoming actions are temporarily held in working memory (WM)[6–8], which is theorized to allow direct access to a limited amount of goal-relevant information[9–11]. In addition to active maintenance, the WM system provides a platform to allow flexible selection of information for upcoming behavior via attention. In all of our experiments so far, participants were able to select and integrate arbitrary,

action features into the conjunctive representations “on the fly”. To form goal-compatible, conjunctive representations for each action, subjects likely used top-down attention to select relevant action features held in WM. In fact, while forming conjunctive representations, subjects spontaneously selected critical action features over other features that are redundant to specify a given action (e.g., cue words for each rule: “vertical” or “updown”). In addition, in an ongoing follow-up study (n = 26), we explicitly manipulated the relevance of action features on a trial-by-trial basis (Fig. 1a). Out of two possible action rules for each block, one rule was randomly selected to be relevant and another one became irrelevant. The relevant rule further specified the target and the distractor stimulus position, which defined the required action for a given trial (Fig. 1a). This paradigm allowed us to test whether people could flexibly select now-relevant features to form the conjunctive representations of the required action over currently irrelevant features. Results showed that subjects were able to selectively attend to the relevant features to form the conjunctive representations of the required action, albeit in an imperfect manner (Fig. 1b). These observations raise the question of underlying neural mechanisms of top-down attentional control over the selection of action features in WM and potential capacity constraints.

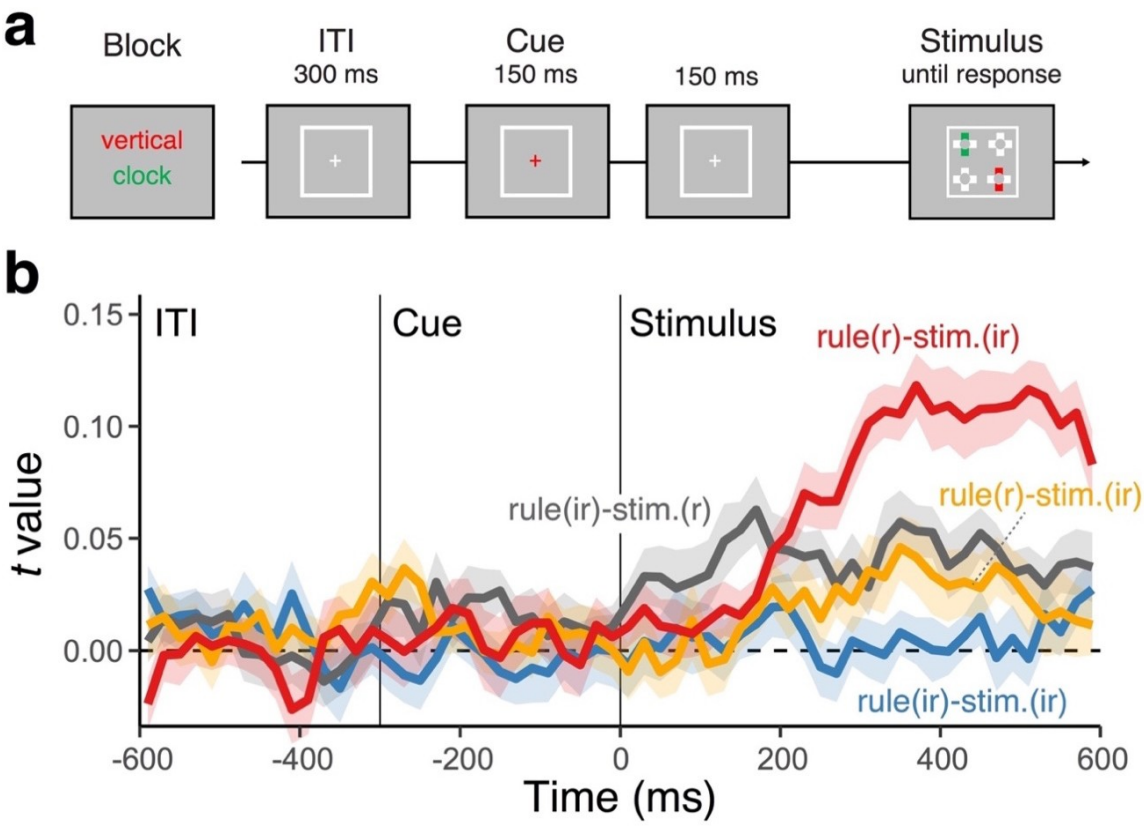


Fig 1. a, Trial events in the rule-selection task that demands attentional selection of the relevant rule/stimulus-feature over the irrelevant ones on a trial-by-trial basis (*n* = 26). Two randomly selected action rules were assigned either a red or green color at the beginning of each block. The relevant rule was randomly cued by the color of the fixation cross for each trial. Subjects applied the rule to the bar in a corresponding color (e.g., use the vertical rule to the red bar location as the starting point). Other procedures are identical to Exp.2 in Chapter II. **b**, Average, single-trial *t*-values for rule-S-R conjunction derived from the RSA analysis (see ***Representational similarity analysis*** in Chapter II). Shaded regions specify the standard error around the mean. Four different conjunctive representations were independently estimated: a target conjunction where both rule and stimulus are relevant (rule(r)-stim.(r), two "misbinding" conjunctions where either rule or stimulus feature is irrelevant (rule(r)-stim.(ir) or rule(ir)-stim.(r): "r" stands for relevant and "ir" denotes irrelevant feature), and a distractor conjunction where both features are irrelevant (rule(ir)-stim.(ir)).

One possibility is that conjunctive representations are maintained in WM by population-level neural ensembles that adjust response properties based on the relevance to the current goal[12,13]. Adaptive coding theory of cognitive control suggests, in prefrontal cortex (PFC), task parameters directly shape the tuning properties of neurons that are not inherently "fixed" to specific features but rather adapt to whatever task-relevant information needs to be represented on an ad hoc basis[14–17]. Such a coding scheme

enables computation in a context-dependent manner, and thereby could form the backbone of WM and cognitive control. Yet, the fact that prefrontal neurons are organized in terms of goal-relevance, rather than in a spatially-homogeneous manner, makes it challenging to detect localized effects[18]. When combined with the population-level neural dynamics, the adaptive coding mechanism allows the system to flexibly preserve goal-relevant information even without sustained, delay activity of individual neurons with fixed tuning profiles[12,19–21]. For example, information could be held or "relayed" by the cascade of firing patterns of unique populations of neurons, or by effective connectivity in the network such as through short-term synaptic plasticity[22,23] and/or synchronous oscillations[24]. One way to test the stability of population-codes is applying decoders trained at one time sample to other time samples, and test if the pattern generalizes over time (i.e., temporal generalization method)[25,26]. The functional properties of stable vs. dynamic population-codes are still activity debated. On the one hand, the stationary responses provide constant mapping between information and the patterns, which establish straightforward readout by downstream neurons. On the other hand, dynamic patterns may code additional information (e.g., elapsed time) related to maintenance or make representations less prone to interference from external inputs[20]. Yet, representations in WM often exhibit dynamic, non-stationary patterns that gradually become stationary[19–21]. Consistently, our results also showed the underlying neural responses encoding conjunctions was temporally dynamic: Different patterns of EEG relayed the conjunctive information for every moment, and the pattern became more stable only later in the trial (Fig. 2). Yet, strikingly, when we focus on the *individual moments* of the pattern, such dynamic and highly idiosyncratic neural responses were consistent within individuals over several months (Supplementary Fig. 9 in Chapter II).

Together, these observations suggest that the dynamic nature of neural responses is inherent to action selection. This further raises the question of how the dynamic population-coding scheme that is common to WM representations contribute to the formation and maintenance of conjunctions.

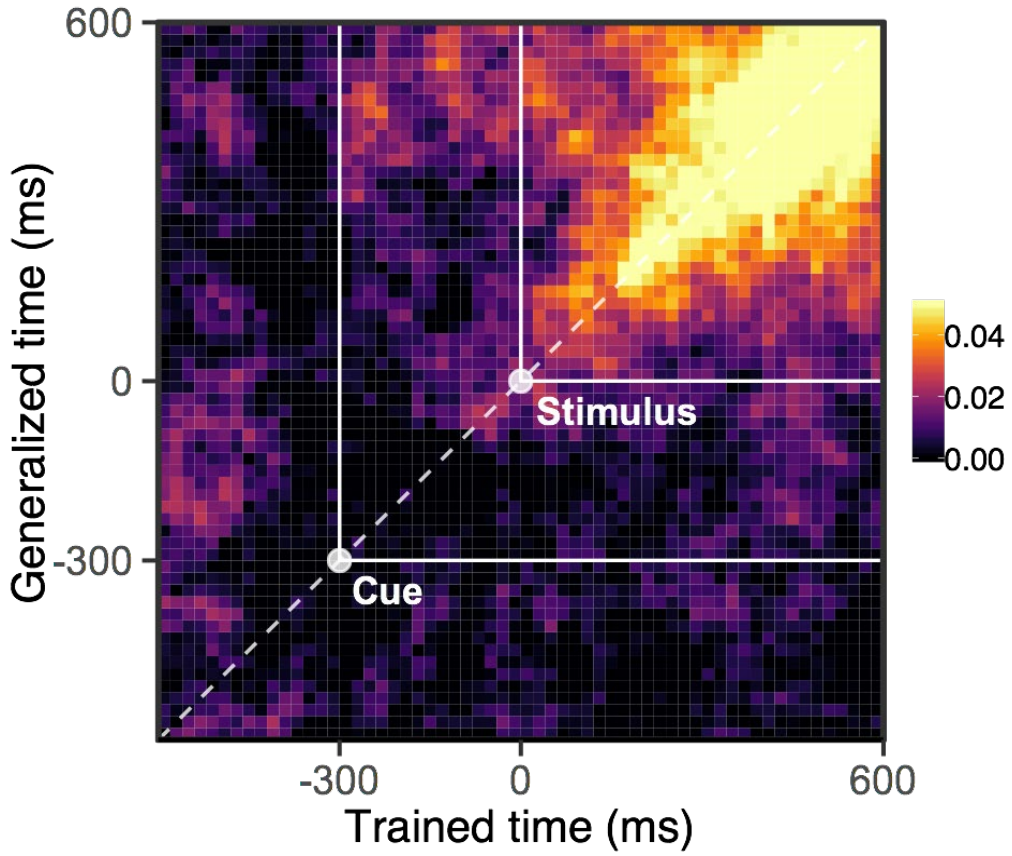


Fig 2. Cross-temporal generalization analysis of the conjunctive representation (using data from Exp.1 in Chapter II). Decoders trained at specific time points were tested on all other time samples in the trial-interval, which generated a matrix of decoding results (Y-axis: time points where decoders were generalized to; X-axis: time points where decoders were trained). Then, the conjunctive representation was estimated via RSA (see ***Representational similarity analysis*** in Chapter II) in all decoding outputs in the matrix. Thus, the height (color) of on-diagonal cells correspond the time-resolved RSA result (Fig. 3a in Chapter II), and the off-diagonal cells show generalization to other time samples. In the early response selection phase, underlying population dynamics is not stationary (i.e., a unique EEG pattern relays information over time), and it gradually becomes more stable over time (i.e., more generalizable to other time samples).

As indicated, maintaining conjunctive representations likely recruits WM and relies on dynamic and goal-contingent population-coding[27,28]. In particular, as further discussed below, the population dynamics of neurons with nonlinear, mixed selectivity

(i.e., tuning to the mixture of task aspects) is proposed to be critical for active maintenance of goal-relevant information[29,30]. Parthasarathy and colleagues (2017) showed that mixed-selectivity neurons generate "morphed" population dynamics in response to a distractor by rapidly adapting their tuning properties. Higher degree of morphing predicted less information loss by a distractor, suggesting representations were sustained at the level of the population-codes while underlying neural responses changed before/after the distractor onset. This property of shifting population dynamics was unique to nonlinear, mixed-selectivity neurons unlike neurons with classical selectivity (i.e., tuning to a particular value of feature like red color). By extending our paradigm, we could test active maintenance of the goal-compatible conjunctive representations against irrelevant information (Fig. 3ab). Subjects would be instructed to prepare and maintain specific actions while ignoring irrelevant distractors. This allows us to investigate how the population ensembles encoding conjunctive representations change to preserve information against a distractor. We predict that the relevant conjunctive code is maintained actively even in the presence of the distractor (and no conjunctions that integrate the distractor information will be formed). However, the conjunctive representation should be preserved in a morphed state that induces a non-generalizable pattern over delay periods (Fig. 3b). Negative effects of distractors on behavior and the conjunctive representations (i.e., loss of information) should be dependent on the degree of morphing of the pattern. Such an experiment would test constraints in active maintenance of conjunctive representations as well as the nature of underlying population codes, which are important to further advance our understanding of flexible action control.

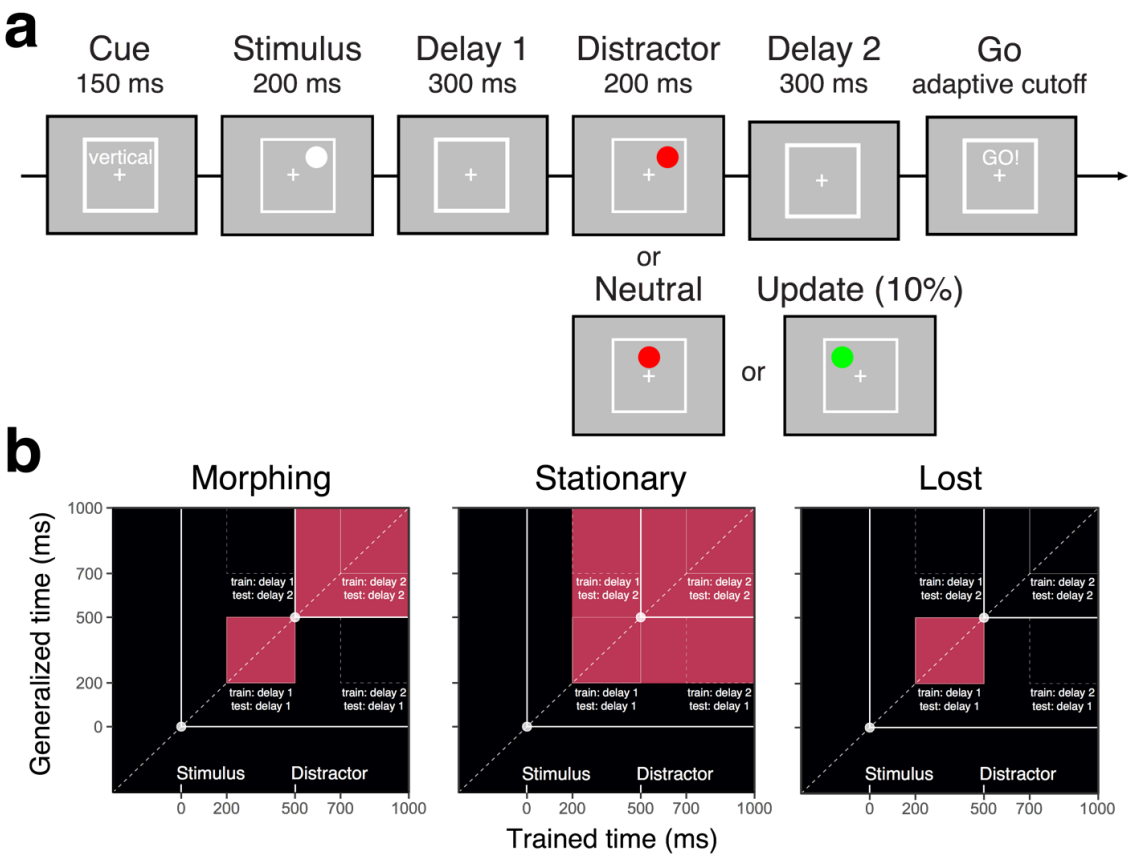


Fig 3 a, Trial events in the rule-based selection task with a distractor during maintenance. Participants will be instructed to prepare and maintain the cued action until "Go" phase (with the adaptive cutoff times to encourage maintenance). During the retention period, three possible events occur: a task-irrelevant distractor appearing at possible target locations ("Distractor" condition), a neutral object appearing at the fixed non-target position ("Neutral" condition), or a new target appearing ("Update" condition) in 10% of trials. Update trials will not be analyzed. Other procedures are identical to the Exp.2 in Chapter II. **b**, Predicted patterns of cross-temporal generalization of conjunctive representations (see also Fig.2). If the distractor interferes with the maintained conjunctive representation, information should become undetectable in the second delay period ("Lost"). Alternatively, if the conjunctive representation can be maintained against a distractor, information should be retained in the second delay either in temporally generalizable manner ("Stationary") or in altered population dynamics ("Morphing").

Mixed selectivity and high-dimensional neural responses

Action-specific conjunctive representations, so far, have been characterized in terms of the representational "content" rather than the "format" of the underlying neural responses. Our analysis extracted nonlinear, neural responses that were unique to the mixture of action-features, over and above other constituent features. Similarly, recent studies have highlighted the importance of neurons with mixed selectivity that show

diverse, nonlinear, and complicated tuning for the combination of multiple task features during goal-directed behavior[30–32]. Mixed selectivity neurons are hypothesized to play a central computational role in cognitive control by shaping the "format" of representations, in particular the dimensionality of neural responses. The dimensionality of information is fundamentally related to linear separability (i.e., readout) of patterns[33], that is the accessibility of information to downstream receiving neurons. Here, I discuss how we might use our approach to get at the computational foundations of the neural responses that are critical for the formation of conjunctive representations.

Computational theories have proposed that the dimensionality of neural responses could control a trade-off between flexibility and efficiency of representations. High-dimensional formats allow the system to flexibly encode diverse combinations of input-output functions in separable patterns of firing. This exponentially expands the number of possible linear readout of information to downstream neurons[31,32,34,35]. In contrast, low-dimensional formats discard information along irrelevant dimensions, thus efficiently coding information that is more generalizable and robust to variance. The flexibility enabled by a high-dimensional format may be particularly important for cognitive control functions, which require representations that reflect arbitrary contingencies of the current task context.

Recent studies of single-neuron electrophysiology in non-human primates showed that the majority of neurons, in particular in frontal and parietal cortices that are known to support rule-guided actions[36], exhibit nonlinear, mixed selectivity to the combination of task aspects (cues, rules, stimuli, and responses)[30,32,37–40]. These neurons show the unique property of upregulating the dimensionality of neural responses at the population-level. In contrast, neurons with classical or *linear* mixed selectivity (i.e., tuning to one or few

instances of features in a linearly additive manner) do not increase the dimensionality effectively. Computationally, the increase of dimensionality is known to make low-dimensional features entailed in the mixed tuning profiles easily separable by downstream receiver neurons. Importantly, high-dimensional activity selectivity diminishes prior to response errors, which parallels the notion of conjunctive representations being critical for successful actions, as event-file theory suggests. A similar coding scheme has been identified in the recurrent networks models with random and sparse connections (sometimes referred to as "reservoir computing"), which resembles the empirical cortical connectivity in PFC[32,35,41]. Artificial neurons that are trained to perform a variety of cognitive tasks in these recurrent networks are also known to spontaneously encode information using mixed selectivity, which generates high-dimensional dynamics at the population-level[42]. Together, the current evidence suggests that mixed selectivity of neurons is the computational basis of conjunctive neural representations: They provide the high-dimensional format of population activity that would implement a substantial number of input-readout functions such that downstream neurons could separate task-relevant information efficiently.

The nonlinear nature of the conjunctive representations implies that mixed selectivity neurons may contribute feature integration and control the dimensionality of neural responses during action selection. Specifically, high-dimensional formats should enhance the strength and readability of conjunctive representations, which could be up/downregulated according to the demand of the current goal. An accurate estimation of the dimensionality in noisy neural responses is challenging. However, recently new methods have been proposed to directly link the dimensionality to behavior in both non-human primates and humans[43–4546,4743–45]. For example, Rigotti et al (2000) suggests that

by computing the number of implementable binary classifications between experimental conditions over noise, one can estimate the dimensionality of neural responses. The neural activity in N units (e.g., neurons or EEG electrodes) could be represented as a point in N-dimensional space for C different experimental conditions, where its dimensionality D equals to the rank of N x C matrix in noiseless cases. When the dimensionality is at maximum, neural responses are linearly independent (i.e., full rank), which allows a linear decoder to read out all possible binary separations (2^C conditions) between conditions perfectly. Critically, the estimated dimensionality scales based on the tuning properties or selectivity of the underlying neural responses. Only if neural responses exhibit nonlinear, mixed selectivity the dimensionality can be higher than the level that other simpler tuning profiles could implement.

If high degrees of dimensionality in neural responses and conjunctive representations actually reflect the same underlying mechanism (i.e., neurons with nonlinear mixed selectivity), then we should find them to be related across conditions and individuals. For example, Fig. 4ab shows an experiment that would contrast action selection in either a low- or a high-dimensional action space. This experiment would allow us to combine the binary-classification method with the time-resolved RSA in order to analyze both dimensionality and the presence of conjunctive representations side by side. Specifically, subjects perform the rule-based selection task (similar to the paradigm for the Exp.1 in Chapter II), where all (4-rules blocks) or one (1-rule blocks) of the action rules becomes relevant for a given block. We can test here whether (a) subjects flexibly adjusts the dimensionality of neural responses according to the changes in the number of relevant cases in the action space (Fig. 4b) and (b) the dimensionality modulates the strength of conjunctive representations (i.e., contents). The maximum

dimensions that spans action constellations (4 rules x 4 stimuli/responses) are 16, which allows (2^16) -2 (i.e., 65534, excluding 2 cases out of 2^16 that assign the same label) possible ways to classify binary cases. If the observed neural responses *only* reflect classical or pure selectivity (e.g., tuning to one instance, such as the vertical rule), the maximum possible dimension reduces to (rule: 4 - 1= **3**) + (stimulus: 4 - 1 = **3**) + (response: 4 - 1 = **3**) + **1** (additional degrees of freedom given by the displacement of the patterns of activity from the origin) = **10**, which leads to 2^10 (1024) binary cases. Based on the computational role of neurons with mixed selectivity, the dimensionality and conjunctive representations should covary in the same direction according to the size of the action space. Specifically, as the number of possible actions increases (e.g., 4-rule blocks), the dimensionality should be scaled up to implement more separable patterns that differentiate unique action scenarios—and also enable conjunctive representations. High dimensional formats should be particularly important to implement higher-order conjunctions (e.g., rule-S-R conjunctions) compared to rule-independent S-R conjunctions, where heightened separability among action constellations is more useful. In addition, the observed, robust relationship between the quality of conjunctive representations and trial-to-trial performance may be mediated by changes in dimensionality. Such an experiment would be an important next step to connect our current findings and theories of action control in humans to mixed selectivity neurons that has been exclusively studied in non-human animals, which enables systematic comparisons across species and between computational models.

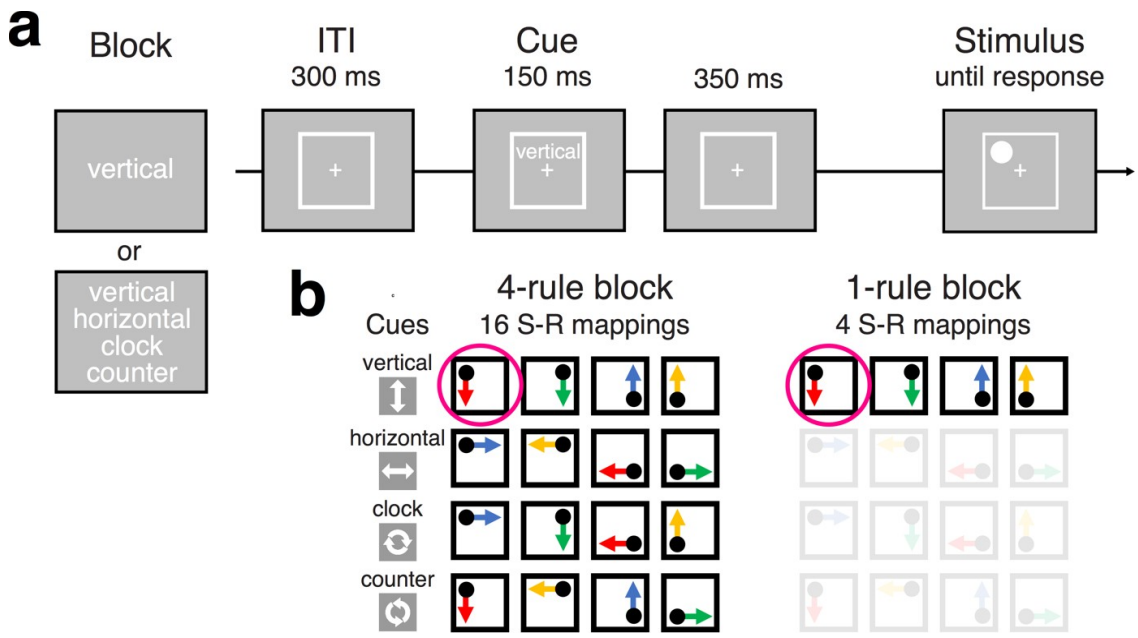


Fig 4 a, Trial events in the rule-based selection task with the reduced action space by reducing a number of relevant action rules. Across blocks, the full or reduced action space changes the number of relevant action constellations (the relevant rule is randomly selected in 1-rule blocks). Other procedures are identical to the Exp.2 in Chapter II. **b**, Schematic diagrams of the space of action constellations in 4-rule blocks and 1-rule blocks. The same action (e.g., a vertical translation of a dot at the left-top corner) occur in different action contexts with different number of potential scenarios that need to be distinguished during action selection.

Effect integration and representation learning

Actions are executed to produce specific, goal-compatible consequences. Action features should be weighted more heavily if they are compatible with an anticipated goal (e.g., how you hold chopsticks is less important to taste food, but it becomes relevant to eat fast). In this sense, voluntary action control relies on the knowledge of causal structures (i.e., what factors lead to the intended action outcomes) of the task at hand. Yet, computationally, learning the values of all possible associations between actions, outcomes, and goals would be daunting and slow[48,49]. The complex, multidimensional nature of the world necessitates us to learn how to parse and choose relevant action features in resource-rational manner[50]. Modern theories of reinforcement learning suggest that the integration of action outcomes depends on building the knowledge of structures

in the task space—learning how agents should cluster observations and construct efficient task representations to guide actions[51–55]—termed as *representation learning*. By evaluating action outcomes people can infer the causal structures and guide selective attention during action control, which ultimately constrains how we generalize actions in future. The fact that our results indicate that action selections rely on conjunctive representations raises the question: how such integrated representations contribute to the integration of action outcomes and how they change over the course of learning the causal structures of a task?

In our paradigm so far, action outcomes were not distinguishable from the states associated with making responses (e.g., the somatosensory experience associated with a particular button press). Therefore, action consequences were always stable and predictable as long as actions were accurate. In this sense, how conjunctive representations reflect the anticipated outcomes is unclear, which leaves the intriguing possibility that action outcomes (rather than just responses as outcomes) were integrated as a part of conjunctive representations. Indeed, some theories of action control explicitly suggest that voluntary actions are triggered by automatically activating anticipated consequences, which could become a part of the event file[56–58]. By observing the co-occurrence of action (e.g., turning on a room light) and outcome (i.e., a room gets brighter), the consequence of actions become integrated and automatically activated during action selection. For instance, Kunde (2001) demonstrated that after people have been repeatedly exposed to arbitrary events (e.g., a loud or soft tone) with independent S-R selection (e.g., pressing a response button strongly or softly based on a color of stimulus), the compatibility between the outcomes (a loud vs. soft tone) and the irrelevant responses (strong vs. soft button pressing) induces cross-modal compatibility costs. This

suggests people have an automatic tendency to use action outcomes to guide selection. Yet, this poses a fundamental challenge: how do agents know what action representations to assign the credit of action outcomes? As in our studies showing that both constituent features and these conjunctions were all predictive of performance (Fig. 3 and 7 in Chapter II, Fig. 5 in Chapter III and Fig 1 in Chapter IV), any everyday goal-directed actions are likely to be guided by multiple action representations. If agents do not distinguish between the causal representations and the simply available/activated representations (e.g., attributing the taste of food to the way you hold chopsticks), resulting incorrect credit assignment[59,60] would lead to suboptimal action control in future.

To assign the values of actions appropriately, agents must learn the connections between experiences and outcomes. Traditionally, causal learning has been studied in variants of associative learning paradigms[61,62]. Using a variety of discrimination tasks, these studies confirmed that, when confronted with multiple features in the environment, humans could flexibly solve associative learning problems in both an elemental (attaching the values to each variable individually) or a configural manner (assigning the values to the specific combination of variables selectively). Numerous factors (e.g., prior knowledge and experiment instruction) that influence the choice of strategies were identified[63,64]. In functional neuroimaging studies, the hippocampus was found to play a critical role in representing information in separable "units" that values could be assigned to (in particular for configural learning), which is consistent with its theorized computational role in pattern separation in memory formation[65–69]. In addition, recent advancements in theories of reinforcement learning also highlighted the importance of structure learning and attention in causal inference[51,52,54,70,71]. When humans integrate (reward) outcomes to guide actions, we spontaneously form efficient, low-dimensional,

latent structures that represent causal associations in clusters, rather than in individual links between stimuli and feedback, as task-states representations. Efficient task representations, in turn, enable us to selectively attend to determinant features that cause action outcomes, which predicate future generalization of learned actions in new situations by facilitating attentional selection of features based on the hypothesized structure of the task space. One example of such a process is the use of feature-based attention as a heuristic solution during reinforcement learning[72,73]. In dynamic multi-dimensional environments, particularly during early learning phase, humans show a strong bias to assign credits to multiple generalizable features (i.e., colors) in parallel, then construct values of actions by combining these features. A configural solution is used only after a substantial amount of learning has taken place in low-dimensional environments[72]. These studies highlight the fact that we can flexibly adjust the strategies in integration of action outcomes, reflecting the knowledge of higher-order structures of the task space.

Our results so far suggest that both constituent and conjunctive representations of action features are highly active concurrently during action selection. One possible function of conjunctive representations could be to provide an efficient framework to integrate the outcome of actions to build the causal structure of the tasks. The conjunctive representations integrate a minimal set of attended, goal-relevant features, which are sufficient to elicit (or cause) the planned actions (see Fig. 3 and Fig. 7 in Chapter I, Fig. 5 in Chapter II, and Fig.1 in this chapter).

Encoding a complete episode of actions into integrated representations may reduce the number of targeted neural activity to efficiently assign credits of action outcomes. In other words, conjunctive representations serve as a starting point for post-

action evaluation to build (and reduce if possible) the latent causal structures of actions. Whenever the causal structures of actions can be reduced to the constituent features (e.g., a missed shot in soccer due to poor leg control, when no other factors, such as ground conditions indeed mattered), low-dimensional representations (e.g., a motor representation for using a leg) could be *extracted* and *reactivated* from the complete, conjunctive representation. After learning, effects of actions in turn become integrated to the conjunctions of other determinant action features (e.g., rule-S-R conjunctions in our experiments), which forms a higher-level conjunctive representation. Hierarchically organized conjunctive representations (e.g., S-R conjunctions) may exist to link different contingencies in action outcomes for future generalization that require different degrees of abstraction (Fig .5 in Chapter II).

To test these notions, we need a paradigm that allows us 1) to model learning of latent causal structures in a multidimensional feature space, 2) to separate processing for action selection and outcome evaluation (e.g., reactivation), and 3) to decode determinant action features and the conjunction (Fig. 5). Subjects will perform a variant of the rule-based selection task where different action features (or conjunctions) consistently lead to different outcomes. Because the value of the first action partially depends on the quality of the subsequent action, subjects are encouraged to anticipate the effect of the original action. The causal structures of action outcomes are mapped onto the entire action space, which could be inferred via elemental or configural, feature-based attention (Fig. 5b). In the elemental condition, the use of the specific constituent feature (e.g., a vertical rule) increases the probability of a specific action outcome (e.g., a fumble ball at the right top corner, which requires a quick right-top key press). Thus, the latent causal structures correspond to one of the constituent features dimensions (i.e., rule,

stimulus, or response-dimension). In contrast, in the configural condition, subjects cannot reduce the causal structure to the level of one constituent feature, which would encourage them to learn action outcomes for each action scenario individually (i.e., based on conjunctions). During learning, we predict subjects form the minimal latent causal structures that link the executed actions and the observed effects. Learning should be more efficient in elemental than configural causal structures[70,71]. The second-order responses (i.e., fumble detection) should improve as action outcomes of the first responses become more anticipated. Among multiple candidate models that explained learning of values in a multi-dimensional environment, a class of models that capture attentional selection that weights specific feature-dimensions should explain the variability in the second-order responses in the elemental condition. In addition, the conjunctive representation of the original action should be reactivated (or maintained) during the outcome presentation, which in turn predicts reactivation of the constituent feature corresponding to the latent causal structure. This relationship should be positively correlated with attentional weights estimated from performance in the second-order responses. After learning, subjects should begin to integrate the anticipated effects as a part of the conjunctive representations during selection of the first response even before the presentation of outcomes. Taken together, our studies so far demonstrated that high-dimensional, conjunctive representations are formed for specific actions, which contain a complete profile of to-be-executed actions, yet do not easily generalize to other situations. Understanding the role of conjunctive representations in integration of action outcomes could shed light on how we balance generalizability and specificity of representations during action control.

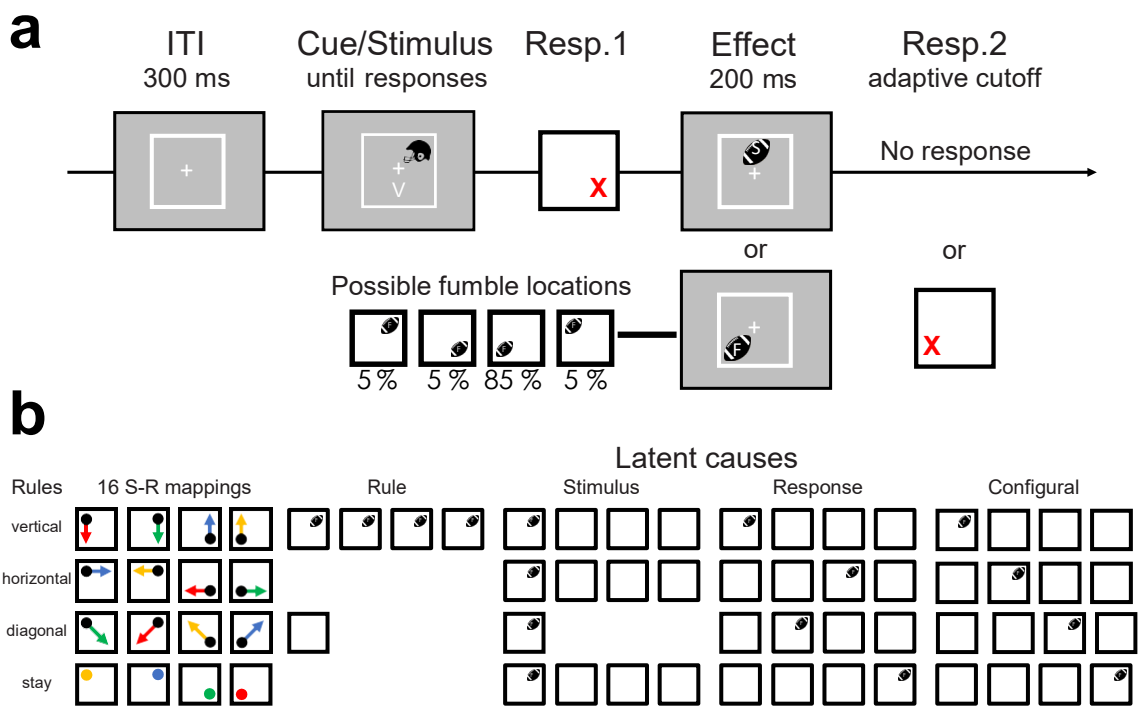


Fig 5 a, Trial events in the rule-based selection task that promotes the integration of action outcomes on the basis of a causal structures. Participants will be instructed a cover story of being a player in the Oregon Ducks football team. For every trial, the combination of a movement tactic and a starting position specifies the required action (e.g., run vertically from let-top corner), which is identical to the task in previous experiments, where the conjunctive representations were observed. The correct response leads two possible outcomes: the "touchdown" condition (denoted by a football with a letter S) where no further responses are required (i.e., correct rejection) or the "fumble" condition (denoted by a football with a latter F) which shows the football at one of the four corners based on the assigned probabilities. In the fumble condition, participants need to press the spatially compatible key to the fumbled football (i.e., hit). Scores in the touchdown condition and the "recovered" plays in the fumble condition will be added to the total incentives. Other procedures are identical to Exp.1 in Chapter II. **b**, Possible causal structures to be learned that are mapped on to the entire action constellations. Actions denoted by the football will lead to the fumble condition (note the football could appear at any of the four corners). In the rule-, stimulus-, response-feature condition (i.e., the elemental condition), latent causal structures directly correspond to one of the constituent action features. In the configural condition (note one example of the configural condition is shown), the causal structures cannot be reduced to lower dimensions, and thus, learned outcomes (values) of each action cannot be generalized along the feature dimensions.

Conclusion

The present project was aimed at characterizing dynamic control during action selection, focusing on the formation of conjunctive representations that bind critical low-dimensional task components into the "exact" control representation for individual

actions. A data-driven, time-resolved decoding approach provided unbiased trajectories of multiple action representations with high temporal resolution and revealed how these representations impacted the variability in behavior at the level of single trials. With these methods, we provided a direct test of an important cognitive theories of action control (event-file theory). This is also the first step towards connecting such theories to results about population-coding dynamics of neurons with nonlinear mixed-selectivity in non-human animals using single-cell electrophysiology. In this final chapter, I have presented a few examples that can extend our work towards a better understanding of the functional role and neural underpinnings of conjunctive representations.

APPENDICES

A. EEG RECORDING AND PREPROCESSING

Electroencephalographic (EEG) activities were recorded from 20 tin electrodes held in place by an elastic cap (Electrocap International) 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 EEG and EOG were amplified with an SA Instrumentation amplifier with a bandpass of 0.01–80 *Hz* and were digitized at 250 *Hz* in LabView 6.1 running on a PC.

B. TIME-FREQUENCY ANALYSIS

Temporal-spectral profiles of single-trial EEG data were obtained via complex wavelet analysis by applying time-frequency analysis to preprocessed EEG data segmented 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/2f,* 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 sample point was defined as the squared magnitude of the convolved signal $Z(\text{real}([z(t)]^2 + \text{imag}[z(t)]^2)$ for instantaneous power.

REFERENCES CITED

Chapter I.

1. Norman, D. A. & Shallice, T. Attention to Action: Willed and Automatic Control of Behavior Technical Report No. 8006. (1980).

2. Engle, R. W., Kane, M. J. & Tuholski, S. W. Individual differences in working memory capacity and what they tell us about controlled attention, general fluid intelligence, and functions of the prefrontal cortex. in *Models of working memory: Mechanisms of active maintenance and executive control , (pp* (ed. Miyake, A.) vol. 506 102–134 (Cambridge University Press, xx, 1999).

3. Miyake, A. *et al.* The Unity and Diversity of Executive Functions and Their Contributions to Complex 'Frontal Lobe' Tasks: A Latent Variable Analysis. *Cogn. Psychol.* **41**, 49–100 (2000).

4. Monsell, S. & Driver, J. *Attention and performance XVIII: Control of cognitive processes*. (MIT Press, 2000).

5. Botvinick, M. M. & Cohen, J. D. The computational and neural basis of cognitive control: charted territory and new frontiers. *Cogn. Sci.* **38**, 1249–1285 (2014).

6. Egner, T. *The Wiley Handbook of Cognitive Control*. (John Wiley & Sons, 2017).

7. Duncan, J. An adaptive coding model of neural function in prefrontal cortex. *Nat. Rev. Neurosci.* **2**, 820–829 (2001).

8. Badre, D., Kayser, A. S. & D'Esposito, M. Frontal cortex and the discovery of abstract action rules. *Neuron* **66**, 315–326 (2010).

9. Collins, A. G. E., Cavanagh, J. F. & Frank, M. J. Human EEG uncovers latent generalizable rule structure during learning. *J. Neurosci.* **34**, 4677–4685 (2014).

10. Collins, A. G. E. & Frank, M. J. Cognitive control over learning: creating, clustering, and generalizing task-set structure. *Psychol. Rev.* **120**, 190–229 (2013).

11. Rogers, R. D. & Monsell, S. Costs of a predictible switch between simple cognitive tasks. *J. Exp. Psychol. Gen.* **124**, 207 (1995).

12. Monsell, S. Task switching. *Trends Cogn. Sci.* **7**, 134–140 (2003).

13. Dreisbach, G., Goschke, T. & Haider, H. The role of task rules and stimulus–response mappings in the task switching paradigm. *Psychol. Res.* **71**, 383–392 (2007).

14. Donders, F. C. On the speed of mental processes. *Acta Psychol.* **30**, 412–431 (1969).

15. Posner, M. I. & Mitchell, R. F. Chronometric analysis of classification. Cooper, R. P. & Shallice, T. Hierarchical schemas and goals in the control of sequential behavior. *Psychological review* vol. 113 887–916; discussion 917–31 (2006).

16. Miller, E. K. & Cohen, J. D. An integrative theory of prefrontal cortex function. *Annu. Rev. Neurosci.* **24**, 167–202 (2001).

17. Fuster, J. M. Anatomy of the Prefrontal Cortex. *The Prefrontal Cortex* 7–58 (2008) doi:10.1016/b978-0-12-373644-4.00002-5.

18. Kleinsorge, T. & Heuer, H. Hierarchical switching in a multi-dimensional task space.
Psychol. Res. **62**, 300–312 (1999).

19. Korb, F. M., Jiang, J., King, J. A. & Egner, T. Hierarchically Organized Medial Frontal Cortex-Basal Ganglia Loops Selectively Control Task- and Response-Selection. *J. Neurosci.* **37**, 7893–7905 (2017).

20. Koechlin, E., Ody, C. & Kouneiher, F. The architecture of cognitive control in the human prefrontal cortex. *Science* **302**, 1181–1185 (2003).

21. Badre, D. & Nee, D. E. Frontal Cortex and the Hierarchical Control of Behavior. *Trends Cogn. Sci.* **22**, 170–188 (2018).

22. Badre, D. & D'Esposito, M. Is the rostro-caudal axis of the frontal lobe hierarchical? *Nat. Rev. Neurosci.* **10**, 659–669 (2009).

23. Hommel, B. Event Files: Evidence for Automatic Integration of Stimulus-Response Episodes. *Vis. cogn.* **5**, 183–216 (1998).

24. Hommel, B. Theory of Event Coding (TEC) V2.0: Representing and controlling perception and action. *Atten. Percept. Psychophys.* (2019) doi:10.3758/s13414-019- 01779-4.

25. Mayr, U. & Bryck, R. L. Sticky rules: integration between abstract rules and specific actions. *J. Exp. Psychol. Learn. Mem. Cogn.* **31**, 337–350 (2005).

26. Schumacher, E. H. & Hazeltine, E. Hierarchical Task Representation: Task Files and Response Selection. *Curr. Dir. Psychol. Sci.* **25**, 449–454 (2016).

27. Denkinger, B. & Koutstaal, W. Perceive-decide-act, perceive-decide-act: how abstract is repetition-related decision learning? *J. Exp. Psychol. Learn. Mem. Cogn.* **35**, 742–756 (2009).

28. Logan, G. D. Toward an instance theory of automatization. *Psychol. Rev.* **95**, 492–527 (1988).

29. Logan, G. D. Repetition priming and automaticity: Common underlying mechanisms? *Cogn. Psychol.* **22**, 1–35 (1990).

30. Rigotti, M. *et al.* The importance of mixed selectivity in complex cognitive tasks. *Nature* **497**, 585–590 (2013).

31. *Psychol. Rev.* **74**, 392–409 (1967).

32. Fusi, S., Miller, E. K. & Rigotti, M. Why neurons mix: high dimensionality for higher cognition. *Curr. Opin. Neurobiol.* **37**, 66–74 (2016).

33. Warden, M. R. & Miller, E. K. Task-dependent changes in short-term memory in the prefrontal cortex. *J. Neurosci.* **30**, 15801–15810 (2010).

34. Enel, P., Procyk, E., Quilodran, R. & Dominey, P. F. Reservoir Computing Properties of Neural Dynamics in Prefrontal Cortex. *PLoS Comput. Biol.* **12**, e1004967 (2016).

35. Parthasarathy, A. *et al.* Mixed selectivity morphs population codes in prefrontal cortex. *Nat. Neurosci.* **20**, 1770–1779 (2017).

36. Pashler, H. & Baylis, G. C. Procedural learning: II. Intertrial repetition effects in speeded-choice tasks. *Journal of Experimental Psychology: Learning, Memory, and Cognition* vol. 17 33–48 (1991).

37. Campbell, K. C. & Proctor, R. W. Repetition effects with categorizable stimulus and response sets. *J. Exp. Psychol. Learn. Mem. Cogn.* **19**, 1345–1362 (1993).

38. Verbruggen, F., Logan, G. D., Liefooghe, B. & Vandierendonck, A. Short-term aftereffects of response inhibition: repetition priming or between-trial control adjustments? *J. Exp. Psychol. Hum. Percept. Perform.* **34**, 413–426 (2008).

39. Norman, K. A., Polyn, S. M., Detre, G. J. & Haxby, J. V. Beyond mind-reading: multi- voxel pattern analysis of fMRI data. *Trends Cogn. Sci.* **10**, 424–430 (2006).

40. Stokes, M. G., Wolff, M. J. & Spaak, E. Decoding Rich Spatial Information with High Temporal Resolution. *Trends Cogn. Sci.* **19**, 636–638 (2015).

41. King, J.-R. & Dehaene, S. Characterizing the dynamics of mental representations: the temporal generalization method. *Trends Cogn. Sci.* **18**, 203–210 (2014).

42. Garcia, J. O., Srinivasan, R. & Serences, J. T. Near-real-time feature-selective modulations in human cortex. *Curr. Biol.* **23**, 515–522 (2013).

43. Grootswagers, T., Wardle, S. G. & Carlson, T. A. Decoding Dynamic Brain Patterns from Evoked Responses: A Tutorial on Multivariate Pattern Analysis Applied to Time Series Neuroimaging Data. *J. Cogn. Neurosci.* **29**, 677–697 (2017).

Chapter II.

1. Donders, F.C. On the speed of mental processes. *Acta Psychol.* **30**, 412-431 (1969).

2. Sternberg, S. The discovery of processing stages: Extensions of Donders' method. *Acta Psychol.* **30**, 276-315 (1969).

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

4. Sanders, A.F. & Sanders, A. *Elements of human performance: Reaction processes and attention in human skill* (Psychology Press, 2013).

5. Posner, M.I. & Mitchell, R.F. Chronometric analysis of classification. *Psychol. Rev.* **74**, 392 (1967).

6. Monsell, S. Task switching. *Trends Cogn. Sci.* **7**, 134-140 (2003).

7. Hommel, B., Müsseler, J., Aschersleben, G. & Prinz, W. The theory of event coding (TEC): A framework for perception and action planning. *Behav. Brain Sci.* **24**, 849-878 (2001).

8. Hommel, B. Theory of Event Coding (TEC) V2. 0: Representing and controlling perception and action. *Atten. Percep. Psychophys.*, 1-16 (2019).

9. Schumacher, E.H. & Hazeltine, E. Hierarchical task representation: Task files and response selection. *Curr. Dir. Psychol. Sci.* **25**, 449-454 (2016).

10. Mayr, U. & Bryck, R.L. Sticky rules: integration between abstract rules and specific actions. *J. Exp. Psychol. Learn. Mem. Cogn.* **31**, 337-350 (2005).

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

12. Kleinsorge, T. & Heuer, H. Hierarchical switching in a multi-dimensional task space.
Psychol. Forsch. **62**, 300-312 (1999).

13. Korb, F.M., Jiang, J., King, J.A. & Egner, T. Hierarchically organized medial frontal cortex-basal ganglia loops selectively control task-and response-selection. *J. Neurosci.* **37**, 7893-7905 (2017).

14. Hubbard, J., Kikumoto, A. & Mayr, U. EEG Decoding Reveals the Strength and Temporal Dynamics of Goal-Relevant Representations. *Scientific reports* **9**, 9051 (2019).

15. Garcia, J.O., Srinivasan, R. & Serences, J.T. Near-real-time feature-selective modulations in human cortex. *Curr. Biol.* **23**, 515-522 (2013).

16. Hall-McMaster, S., Muhle-Karbe, P.S., Myers, N.E. & Stokes, M.G. Reward Boosts Neural Coding of Task Rules to Optimize Cognitive Flexibility. *J. Neurosci.* **39**, 8549- 8561 (2019).

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

18. Grootswagers, T., Wardle, S.G. & Carlson, T.A. Decoding dynamic brain patterns from evoked responses: A tutorial on multivariate pattern analysis applied to time series neuroimaging data. *J. Cognit. Neurosci.* (2017).

19. Norman, K. A., Polyn, S. M., Detre, G. J. & Haxby, J. V. Beyond mind-reading: multi- voxel pattern analysis of fMRI data. *Trends Cogn. Sci.* **10**, 424–430 (2006).

20. Colzato, L.S., Van Wouwe, N.C., Lavender, T.J. & Hommel, B. Intelligence and cognitive flexibility: fluid intelligence correlates with feature "unbinding" across perception and action. *Psychon Bull Rev* **13**, 1043-1048 (2006).

21. Collins, A.G. & Frank, M.J. Cognitive control over learning: Creating, clustering, and generalizing task-set structure. *Psychol. Rev.* **120**, 190 (2013).

22. Siegel, M., Buschman, T.J. & Miller, E.K. Cortical information flow during flexible sensorimotor decisions. *Science* **348**, 1352-1355 (2015).

23. Woolgar, A., Jackson, J. & Duncan, J. Coding of visual, auditory, rule, and response information in the brain: 10 years of multivoxel pattern analysis. *J. Cognit. Neurosci.* **28**, 1433-1454 (2016).

24. Rigotti, M.*, et al.* The importance of mixed selectivity in complex cognitive tasks. *Nature*
497, 585 (2013).

25. Marcos, E., Tsujimoto, S., Mattia, M. & Genovesio, A. A Network Activity Reconfiguration Underlies the Transition from Goal to Action. *Cell Rep.* **27**, 2909- 2920.e2904 (2019).

26. Duncan, K., Doll, B.B., Daw, N.D. & Shohamy, D. More Than the Sum of Its Parts: A Role for the Hippocampus in Configural Reinforcement Learning. *Neuron* **98**, 645- 657.e646 (2018).

27. Ballard, I.C., Wagner, A.D. & McClure, S.M. Hippocampal pattern separation supports reinforcement learning. *Nat. Commun.* **10**, 1073 (2019).

28. Bhandari, A., Gagne, C. & Badre, D. Just above chance: is it harder to decode information from prefrontal cortex hemodynamic activity patterns? *Journal of cognitive neuroscience* **30**, 1473-1498 (2018).

29. Schuck, N.W. & Niv, Y. Sequential replay of nonspatial task states in the human hippocampus. *Science* **364**, eaaw5181 (2019).

30. Tang, E.*, et al.* Effective learning is accompanied by high-dimensional and efficient representations of neural activity. *Nat. Neurosci.* **22**, 1000 (2019).

31. Warden, M.R. & Miller, E.K. The representation of multiple objects in prefrontal neuronal delay activity. *Cereb. Cortex* **17**, i41-i50 (2007).

32. Rigotti, M., Ben Dayan Rubin, D.D., Wang, X.-J. & Fusi, S. Internal representation of task rules by recurrent dynamics: the importance of the diversity of neural responses. *Front. Comput. Neurosci.* **4**, 24 (2010).

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

34. Olson, I.R., Page, K., Moore, K.S., Chatterjee, A. & Verfaellie, M. Working memory for conjunctions relies on the medial temporal lobe. *J. Neurosci.* **26**, 4596-4601 (2006).

35. Elsner, B. & Hommel, B. Effect anticipation and action control. *J. Exp. Psychol. Hum. Percept. Perform.* **27**, 229 (2001).

36. Brainard, D.H. The psychophysics toolbox. *Spatial Vision* **10**, 433-436 (1997).

37. Pelli, D.G. The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spat. Vis.* **10**, 437-442 (1997).

38. Cohen, M.X. *Analyzing neural time series data: theory and practice* (MIT Press, 2014).

39. Hastie, T., Buja, A. & Tibshirani, R. Penalized discriminant analysis. *Ann. Stat.*, 73-102 (1995).

40. Stokes, M.G., Wolff, M.J. & Spaak, E. Decoding rich spatial information with high temporal resolution. *Trends Cogn. Sci.* **19**, 636-638 (2015).

41. Stokes, M. & Spaak, E. The Importance of Single-Trial Analyses in Cognitive Neuroscience. *Trends Cogn. Sci.* **20**, 483-486 (2016).

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

43. Mosteller, F. & Tukey, J.W. Handbook of Social Psychology. **2**, 80-203 (1968).

44. Nichols, T.E. & Holmes, A.P. Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Hum. Brain Mapp.* **15**, 1-25 (2002)

45. A. Kikumoto, U. Mayr, Decoding hierarchical control of sequential behavior in oscillatory EEG activity. eLife 7, e38550 (2018).

46. J. J. Foster, D. W. Sutterer, J. T. Serences, E. K. Vogel, E. Awh, The topography of alpha-band activity tracks the content of spatial working memory. J. Neurophysiol. 115, 168-177 (2016).

47. V. Wyart, V. de Gardelle, J. Scholl, C. Summerfield, Rhythmic fluctuations in evidence accumulation during decision making in the human brain. Neuron 76, 847-858 (2012).

48. H. Ruge, S. Jamadar, U. Zimmermann, F. Karayanidis, The many faces of preparatory control in task switching: reviewing a decade of fMRI research. Hum. Brain Mapp. 34, 12-35 (2013).

Chapter III.

1. Hommel, B. Event Files: Evidence for Automatic Integration of Stimulus-Response Episodes. *Vis. cogn.* **5**, 183–216 (1998).

2. Schumacher, E. H. & Hazeltine, E. Hierarchical Task Representation: Task Files and Response Selection. *Curr. Dir. Psychol. Sci.* **25**, 449–454 (2016).

3. Logan, G. D. & Cowan, W. B. On the ability to inhibit thought and action: A theory of an act of control. *Psychol. Rev.* **91**, 295 (1984).

4. Logan, G. D. On the ability to inhibit thought and action: A users' guide to the stop signal paradigm. in *Inhibitory processes in attention, memory, and language , (pp* (ed. Dagenbach, D.) vol. 461 189–239 (Academic Press, xiv, 1994).

5. Swann, N. *et al.* 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 (2009).

6. Wessel, J. R. & Aron, A. R. It's not too late: the onset of the frontocentral P3 indexes successful response inhibition in the stop-signal paradigm. *Psychophysiology* **52**, 472– 480 (2015).

7. Wagner, J., Wessel, J. R., Ghahremani, A. & Aron, A. R. Establishing a Right Frontal Beta Signature for Stopping Action in Scalp EEG: Implications for Testing Inhibitory Control in Other Task Contexts. *J. Cogn. Neurosci.* **30**, 107–118 (2018).

8. Wessel, J. R. β-bursts reveal the trial-to-trial dynamics of movement initiation and cancellation. *J. Neurosci.* (2019) doi:10.1523/JNEUROSCI.1887-19.2019.

9. Aron, A. R., Robbins, T. W. & Poldrack, R. A. Inhibition and the right inferior frontal cortex: one decade on. *Trends Cogn. Sci.* **18**, 177–185 (2014).

10. Coxon, J. P., Stinear, C. M. & Byblow, W. D. Intracortical inhibition during volitional inhibition of prepared action. *J. Neurophysiol.* **95**, 3371–3383 (2006).

11. Duque, J., Greenhouse, I., Labruna, L. & Ivry, R. B. Physiological Markers of Motor Inhibition during Human Behavior. *Trends Neurosci.* **40**, 219–236 (2017).

12. Greenhouse, I., Sias, A., Labruna, L. & Ivry, R. B. Nonspecific Inhibition of the Motor System during Response Preparation. *J. Neurosci.* **35**, 10675–10684 (2015).

13. Labruna, L. *et al.* Generic inhibition of the selected movement and constrained inhibition of nonselected movements during response preparation. *J. Cogn. Neurosci.* **26**, 269–278 (2014).

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

15. Kikumoto, A. & Mayr, U. Conjunctive Representations that Integrate Stimuli, Responses, and Rules are Critical for Action Selection. *bioRxiv* 835652 (2019) doi:10.1101/835652.

16. Mayr, U. & Bryck, R. L. Sticky rules: integration between abstract rules and specific actions. *J. Exp. Psychol. Learn. Mem. Cogn.* **31**, 337–350 (2005).

17. Logan, G. D., Van Zandt, T., Verbruggen, F. & Wagenmakers, E.-J. On the ability to inhibit thought and action: general and special theories of an act of control. *Psychol. Rev.* **121**, 66–95 (2014).

18. Verbruggen, F. *et al.* A consensus guide to capturing the ability to inhibit actions and impulsive behaviors in the stop-signal task. *Elife* **8**, (2019).

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

20. Hommel, B. Theory of Event Coding (TEC) V2.0: Representing and controlling perception and action. *Atten. Percept. Psychophys.* (2019) doi:10.3758/s13414-019- 01779-4.

21. Anderson, M. C. & Hanslmayr, S. Neural mechanisms of motivated forgetting. *Trends Cogn. Sci.* **18**, 279–292 (2014).

22. Castiglione, A., Wagner, J., Anderson, M. & Aron, A. R. Preventing a Thought from Coming to Mind Elicits Increased Right Frontal Beta Just as Stopping Action Does. *Cereb. Cortex* **29**, 2160–2172 (2019).

23. Depue, B. E., Orr, J. M., Smolker, H. R., Naaz, F. & Banich, M. T. The Organization of Right Prefrontal Networks Reveals Common Mechanisms of Inhibitory Regulation Across Cognitive, Emotional, and Motor Processes. *Cereb. Cortex* **26**, 1634–1646 (2016).

24. Guo, Y., Schmitz, T. W., Mur, M., Ferreira, C. S. & Anderson, M. C. A supramodal role of the basal ganglia in memory and motor inhibition: Meta-analytic evidence. *Neuropsychologia* **108**, 117–134 (2018).

25. Anderson, M. C. Neural Systems Underlying the Suppression of Unwanted Memories.
Science vol. 303 232–235 (2004).

26. Anderson, M. C. & Green, C. Suppressing unwanted memories by executive control.
Nature **410**, 366–369 (2001).

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

28. Band, G. P. H., van der Molen, M. W. & Logan, G. D. Horse-race model simulations of the stop-signal procedure. *Acta Psychol.* **112**, 105–142 (2003).

29. Verbruggen, F. & Logan, G. D. Proactive adjustments of response strategies in the stop- signal paradigm. *J. Exp. Psychol. Hum. Percept. Perform.* **35**, 835–854 (2009).

30. Aron, A. R. From reactive to proactive and selective control: developing a richer model for stopping inappropriate responses. *Biol. Psychiatry* **69**, e55–68 (2011).

31. Zandbelt, B. B., Bloemendaal, M., Neggers, S. F. W., Kahn, R. S. & Vink, M. Expectations and violations: delineating the neural network of proactive inhibitory control. *Hum. Brain Mapp.* **34**, 2015–2024 (2013).

32. Vink, M., Kaldewaij, R., Zandbelt, B. B., Pas, P. & du Plessis, S. The role of stop-signal probability and expectation in proactive inhibition. *Eur. J. Neurosci.* **41**, 1086–1094 (2015).

33. Cohen, M. X. *Analyzing Neural Time Series Data: Theory and Practice*. (MIT Press, 2014).

34. Kuhn, M. & Others. Building predictive models in R using the caret package. *J. Stat. Softw.* **28**, 1–26 (2008).

35. Nichols, T. E. & Holmes, A. P. Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Hum. Brain Mapp.* **15**, 1–25 (2002).

Chapter IV.

1. Logan, G. D. Toward an instance theory of automatization. *Psychol. Rev.* **95**, 492–527 (1988).

2. Logan, G. D. Repetition priming and automaticity: Common underlying mechanisms? *Cogn. Psychol.* **22**, 1–35 (1990).

3. Hommel, B. Event Files: Evidence for Automatic Integration of Stimulus-Response Episodes. *Vis. cogn.* **5**, 183–216 (1998).

4. Hommel, B., Müsseler, J., Aschersleben, G. & Prinz, W. The Theory of Event Coding (TEC): a framework for perception and action planning. *Behav. Brain Sci.* **24**, 849–78; discussion 878–937 (2001).

5. Schumacher, E. H. & Hazeltine, E. Hierarchical Task Representation: Task Files and Response Selection. *Curr. Dir. Psychol. Sci.* **25**, 449–454 (2016).

6. Engle, R. W. & Kane, M. J. Executive attention, working memory capacity, and a two- factor theory of cognitive control. *Psychol. Learn. Motiv.* **44**, 145–200 (2004).

7. Miller, E. K. & Cohen, J. D. An integrative theory of prefrontal cortex function. *Annu. Rev. Neurosci.* **24**, 167–202 (2001).

8. Kikumoto, A. & Mayr, U. The Nature of Task Set Representations in Working Memory. *J. Cogn. Neurosci.* **29**, 1950–1961 (2017).

9. Unsworth, N., Fukuda, K., Awh, E. & Vogel, E. K. Working memory and fluid intelligence: capacity, attention control, and secondary memory retrieval. *Cogn. Psychol.* **71**, 1–26 (2014).

10. Drew, T. & Vogel, E. K. Working Memory: Capacity Limitations. *Encyclopedia of Neuroscience* 523–531 (2009) doi:10.1016/b978-008045046-9.00428-9.

11. Oberauer, K. *et al.* Benchmarks for models of short-term and working memory. *Psychol. Bull.* **144**, 885–958 (2018).

12. Stokes, M. G. *et al.* Dynamic coding for cognitive control in prefrontal cortex. *Neuron* **78**, 364–375 (2013).

13. Stokes, M. G., Buschman, T. J. & Miller, E. K. Dynamic Coding for Flexible Cognitive Control. *The Wiley Handbook of Cognitive Control* 221–241 (2017) doi:10.1002/9781118920497.ch13.

14. Duncan, J. An adaptive coding model of neural function in prefrontal cortex. *Nat. Rev. Neurosci.* **2**, 820–829 (2001).

15. Duncan, J. & Miller, E. K. Cognitive focus through adaptive neural coding in the primate prefrontal cortex. in *Principles of frontal lobe function , (pp* (ed. Stuss, D. T.) vol. 616 278–291 (Oxford University Press, xxi, 2002).

16. Meyers, E. M., Freedman, D. J., Kreiman, G., Miller, E. K. & Poggio, T. Dynamic population coding of category information in inferior temporal and prefrontal cortex. *J. Neurophysiol.* **100**, 1407–1419 (2008).

17. Hussar, C. R. & Pasternak, T. Flexibility of sensory representations in prefrontal cortex depends on cell type. *Neuron* **64**, 730–743 (2009).

18. Bhandari, A., Gagne, C. & Badre, D. Just above Chance: Is It Harder to Decode Information from Prefrontal Cortex Hemodynamic Activity Patterns? *J. Cogn. Neurosci.* **30**, 1473–1498 (2018).

19. Watanabe, K. & Funahashi, S. Neural mechanisms of dual-task interference and cognitive capacity limitation in the prefrontal cortex. *Nat. Neurosci.* **17**, 601–611 (2014).

20. Sreenivasan, K. K. & D'Esposito, M. The what, where and how of delay activity. *Nat. Rev. Neurosci.* **20**, 466–481 (2019).

21. Murray, J. D. *et al.* Stable population coding for working memory coexists with heterogeneous neural dynamics in prefrontal cortex. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 394–399 (2017).

22. Mongillo, G., Barak, O. & Tsodyks, M. Synaptic theory of working memory. *Science*
319, 1543–1546 (2008).

23. Mongillo, G., Rumpel, S. & Loewenstein, Y. Intrinsic volatility of synaptic connections—a challenge to the synaptic trace theory of memory. *Curr. Opin. Neurobiol.* **46**, 7–13 (2017).

24. Fries, P. A mechanism for cognitive dynamics: neuronal communication through neuronal coherence. *Trends Cogn. Sci.* **9**, 474–480 (2005).

25. King, J.-R. & Dehaene, S. Characterizing the dynamics of mental representations: the temporal generalization method. *Trends Cogn. Sci.* **18**, 203–210 (2014).

26. Marti, S., King, J.-R. & Dehaene, S. Time-Resolved Decoding of Two Processing Chains during Dual-Task Interference. *Neuron* **88**, 1297–1307 (2015).

27. Romo, R., Brody, C. D., Hernández, A. & Lemus, L. Neuronal correlates of parametric working memory in the prefrontal cortex. *Nature* vol. 399 470–473 (1999).

28. Miller, E. K., Erickson, C. A. & Desimone, R. Neural mechanisms of visual working memory in prefrontal cortex of the macaque. *J. Neurosci.* **16**, 5154–5167 (1996).

29. Parthasarathy, A. *et al.* Mixed selectivity morphs population codes in prefrontal cortex.
Nat. Neurosci. **20**, 1770–1779 (2017).

30. Rigotti, M., Ben Dayan Rubin, D., Wang, X.-J. & Fusi, S. Internal representation of task rules by recurrent dynamics: the importance of the diversity of neural responses. *Front. Comput. Neurosci.* **4**, 24 (2010).

31. Fusi, S., Miller, E. K. & Rigotti, M. Why neurons mix: high dimensionality for higher cognition. *Curr. Opin. Neurobiol.* **37**, 66–74 (2016).

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

33. Vapnik, V. The Support Vector Method of Function Estimation. in *Nonlinear Modeling: Advanced Black-Box Techniques* (eds. Suykens, J. A. K. & Vandewalle, J.) 55–85 (Springer US, 1998).

34. Asaad, W. F., Rainer, G. & Miller, E. K. Task-specific neural activity in the primate prefrontal cortex. *J. Neurophysiol.* **84**, 451–459 (2000).

35. Enel, P., Procyk, E., Quilodran, R. & Dominey, P. F. Reservoir Computing Properties of Neural Dynamics in Prefrontal Cortex. *PLoS Comput. Biol.* **12**, e1004967 (2016).

36. Siegel, M., Buschman, T. J. & Miller, E. K. Cortical information flow during flexible sensorimotor decisions. *Science* **348**, 1352–1355 (2015).

37. Warden, M. R. & Miller, E. K. Task-dependent changes in short-term memory in the prefrontal cortex. *J. Neurosci.* **30**, 15801–15810 (2010).

38. Jun, J. K. *et al.* Heterogenous Population Coding of a Short-Term Memory and Decision Task. *Journal of Neuroscience* vol. 30 916–929 (2010).

39. Zhang, C. Y. *et al.* Partially Mixed Selectivity in Human Posterior Parietal Association Cortex. *Neuron* **95**, 697–708.e4 (2017).

40. Churchland, M. M. & Shenoy, K. V. Temporal complexity and heterogeneity of single- neuron activity in premotor and motor cortex. *J. Neurophysiol.* **97**, 4235–4257 (2007).

41. Pascanu, R. & Jaeger, H. A neurodynamical model for working memory. *Neural Netw.* **24**, 199–207 (2011).

42. Yang, G. R., Joglekar, M. R., Song, H. F., Newsome, W. T. & Wang, X.-J. Task representations in neural networks trained to perform many cognitive tasks. *Nat. Neurosci.* **22**, 297–306 (2019).

43. Kobak, D. *et al.* Demixed principal component analysis of neural population data. *eLife Sciences* **5**, e10989 (2016).

44. Rigotti, M. *et al.* The importance of mixed selectivity in complex cognitive tasks. *Nature* **497**, 585–590 (2013).

45. Bhandari, A., Benna, M., Rigotti, M., Fusi, S. & Badre, D. Measuring prefrontal representational geometry: fMRI adaptation vs pattern analysis. *2019 Conference on Cognitive Computational Neuroscience* (2019) doi:10.32470/ccn.2019.1162-0.

46. Tang, E. *et al.* Effective learning is accompanied by high-dimensional and efficient representations of neural activity. *Nat. Neurosci.* 1 (2019).

47. Ahlheim, C. & Love, B. C. Estimating the functional dimensionality of neural representations. *Neuroimage* **179**, 51–62 (2018).

48. Sutton, R. S. & Barto, A. G. Reinforcement learning: an introduction MIT Press. *Cambridge, MA* (1998).

49. Barto, A. G. & Mahadevan, S. Recent Advances in Hierarchical Reinforcement Learning. *Discrete Event Dyn. Syst.: Theory Appl.* **13**, 41–77 (2003).

50. Griffiths, T. L., Lieder, F. & Goodman, N. D. Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. Topics in Cognitive Science, 7 (2), 217-229. (2015).

51. Gershman, S. J. & Niv, Y. Learning latent structure: carving nature at its joints. *Curr. Opin. Neurobiol.* **20**, 251–256 (2010).

52. Radulescu, A., Niv, Y. & Ballard, I. Holistic Reinforcement Learning: The Role of Structure and Attention. *Trends Cogn. Sci.* **23**, 278–292 (2019).

53. Niv, Y. Learning task-state representations. *Nat. Neurosci.* **22**, 1544–1553 (2019).

54. Collins, A. G. E. & Frank, M. J. Cognitive control over learning: creating, clustering, and generalizing task-set structure. *Psychol. Rev.* **120**, 190–229 (2013).

55. Franklin, N. T. & Frank, M. J. Compositional clustering in task structure learning. *PLoS Comput. Biol.* **14**, e1006116 (2018).

56. Elsner, B. & Hommel, B. Effect anticipation and action control. *J. Exp. Psychol. Hum. Percept. Perform.* **27**, 229–240 (2001).

57. Kunde, W. Response-effect compatibility in manual choice reaction tasks. *J. Exp. Psychol. Hum. Percept. Perform.* **27**, 387–394 (2001).

58. James, W. The principles of psychology / by William James. (1910) doi:10.5962/bhl.title.47583.

59. Akaishi, R., Kolling, N., Brown, J. W. & Rushworth, M. Neural Mechanisms of Credit Assignment in a Multicue Environment. *J. Neurosci.* **36**, 1096–1112 (2016).

60. Shahar, N. *et al.* Credit assignment to state-independent task representations and its relationship with model-based decision making. *Proc. Natl. Acad. Sci. U. S. A.* (2019) doi:10.1073/pnas.1821647116.

61. Melchers, K. G., Shanks, D. R. & Lachnit, H. Stimulus coding in human associative learning: flexible representations of parts and wholes. *Behav. Processes* **77**, 413–27; discussion 451–3 (2008).

62. Pearce, J. M. & Bouton, M. E. Theories of associative learning in animals. *Annu. Rev. Psychol.* **52**, 111–139 (2001).

63. Shanks, D. R. Forward and Backward Blocking in Human Contingency Judgement. *The Quarterly Journal of Experimental Psychology Section B* **37**, 1–21 (1985).

64. Melchers, K. G., Lachnit, H. & Shanks, D. R. Past experience influences the processing of stimulus compounds in human Pavlovian conditioning. *Learn. Motiv.* **35**, 167–188 (2004).

65. Duncan, K., Doll, B. B., Daw, N. D. & Shohamy, D. More Than the Sum of Its Parts: A Role for the Hippocampus in Configural Reinforcement Learning. *Neuron* **98**, 645– 657.e6 (2018).

66. Ballard, I. C., Wagner, A. D. & McClure, S. M. Hippocampal pattern separation supports reinforcement learning. *Nat. Commun.* **10**, 1073 (2019).

67. Yassa, M. A. & Stark, C. E. L. Pattern separation in the hippocampus. *Trends Neurosci.* **34**, 515–525 (2011).

68. Keresztes, A. *et al.* Hippocampal maturity promotes memory distinctiveness in childhood and adolescence. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 9212–9217 (2017).

69. Bakker, A., Kirwan, C. B., Miller, M. & Stark, C. E. L. Pattern separation in the human hippocampal CA3 and dentate gyrus. *Science* **319**, 1640–1642 (2008).

70. Collins, A. G. E. & Frank, M. J. How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis. *Eur. J. Neurosci.* **35**, 1024–1035 (2012).

71. Gershman, S. J., Norman, K. A. & Niv, Y. Discovering latent causes in reinforcement learning. *Current Opinion in Behavioral Sciences* **5**, 43–50 (2015).

72. Farashahi, S., Rowe, K., Aslami, Z., Lee, D. & Soltani, A. Feature-based learning improves adaptability without compromising precision. *Nat. Commun.* **8**, 1768 (2017).

73. Niv, Y. *et al.* Reinforcement learning in multidimensional environments relies on attention mechanisms. *J. Neurosci.* **35**, 8145–8157 (2015).