HHS Public Access

Author manuscript

Trends Cogn Sci. Author manuscript; available in PMC 2022 July 01.

Published in final edited form as:

Trends Cogn Sci. 2021 July ; 25(7): 622–638. doi:10.1016/j.tics.2021.03.011.

Neural coding of cognitive control: The representational similarity analysis approach

Michael C. Freund[1], **Joset A. Etzel**[1], **Todd S. Braver**[1,2,3]

[1]Department of Psychological and Brain Sciences, Washington University in St. Louis, St. Louis, Missouri, USA 63130

[2]Department of Radiology, Washington University in St. Louis, School of Medicine, St. Louis, Missouri, USA 63110

[3]Department of Neuroscience, Washington University in St. Louis, School of Medicine, St. Louis, Missouri, USA 63110

Abstract

Cognitive control relies on distributed and potentially high-dimensional frontoparietal task representations. Yet, the classical cognitive neuroscience approach in this domain has focused on aggregating and contrasting neural measures — either via univariate or multivariate methods — along highly abstracted, one-dimensional factors (e.g., Stroop congruency). Here, we present representational similarity analysis (RSA) as a complementary approach that can powerfully inform representational components of cognitive control theories. We review several exemplary uses of RSA in this regard. We further show that most classical paradigms, given their factorial structure, can be optimized for RSA with minimal modification. Our aim is to illustrate how RSA can be incorporated into cognitive control investigations to shed new light on old questions.

Keywords

executive function; multivariate pattern analysis (MVPA); representational similarity analysis (RSA); prefrontal cortex (PFC); anterior cingulate cortex (ACC)

Towards modeling cognitive control representations

A healthy human mind can set itself towards achieving a goal. This capacity for **cognitive control** (see Glossary) seems to be a central part of what it means to be human: it putatively underlies abilities that are most elaborate in our species [1-3], yet which go characteristically awry within prevalent mental health disorders [1,4-6]. Propelled by this notion, cognitive scientists have devoted more than a half-century of collective effort towards understanding

Correspondence: tbraver@wustl.edu (T.S. Braver).

We have no known conflict of interest to disclose.

how control arises in mind and brain. This understanding has typically been sought through the lens of two complementary cognitive constructs: **representations** and **processes**. A "representation" is a description of the information that the activity of a neural unit (e.g., neuron, ensemble, area) encodes [7,8] (Figure 1, Key Figure; Theory, blue components). Conversely, a "process" is a description of the function of a neural unit abstracted over particulars of the information encoded ([9]; Figure 1, Theory, orange components). Analogously, a black-box computing function is not defined by a set of internal state values, but instead by a general input–output mapping. Many fundamental problems in cognitive control reflect our lack of understanding control representations [10]. Studying these representations, however, has been notoriously elusive: classically, their neural markers have only been indirectly observable, through process-level measures.

In this Review, we suggest that progress in understanding the mechanisms of control has arrived in the form of an expansion in experimental approach: from the classical, which focuses on measuring control processes, to the representational, which explicitly models control representations. This expansion was precipitated by the development of neural population-level analytic techniques, including **multivariate pattern analysis (MVPA)** [11-13] and dimensionality reduction methods [14], which allow neural coding of an unprecedentedly wide range of variables to be measured at the macroscopic scales of functional neuroimaging. But more specifically, we suggest that the MVPA technique of representational similarity analysis (RSA) [15-17] — although it has currently only been sparsely used within cognitive control research — is particularly well-suited for investigation within this domain. This suitability stems from RSA's flexibility in implementation, and as the name suggests, its explicit focus on modeling representations. First, we discuss how the RSA framework can complement the classical approach. We then illustrate the usefulness of RSA more concretely, by reviewing several recent RSA studies that addressed long-standing issues in cognitive control.

The 'classical' approach measures control processes

Classically, cognitive control investigators have designed and analyzed empirical studies using a particular style of experimental psychology. Despite salient differences, most prototypical cognitive control tasks (e.g., [18-28]) share a key design element: an abstract experimental factor that places differential demands on controlled processing. For example, in the color-word Stroop task, the key factor is congruency: whether the task-relevant dimension (hue) is congruent or incongruent with a spatially overlapping, but irrelevant dimension (word; Figure 1, Design). These factors are abstract in the sense that they contain a small number of levels, which collapse across a diverse set of other task-relevant components (e.g., stimuli, rules, response modalities). During analysis, investigators attempt to isolate similarly abstract control processes (e.g., response conflict resolution), by aggregating and contrasting measures along these factors (e.g., Figure 1, Analysis, top).

We refer to this style as the "classical" approach to studying cognitive control: through the scalar lens of process-level contrasts. As a thermometer indicates the temperature, estimates from these one-dimensional contrasts are assumed to indicate the magnitude or efficiency of the control process (i.e., ranging from "high control" to "low control"). Modulations in these

indicators — for example, as a function of task parameters or individual difference variables — are then used to shed light on characteristics of the underlying control processes (e.g., boundary conditions [29,30] or relations with other constructs [31,32]; Figure 1, large orange arrow linking Measures to Theory). The classical approach is foundational (e.g., [29,30,33]; see also [34]) and has undoubtedly shaped the kind of information that investigators seek to learn from associated cognitive control paradigms.

With the advent of cognitive neuroscience, control processes have been mapped to the recruitment of particular areas and networks of the brain. To accomplish this, the classical approach was imported into functional neuroimaging studies, under the predominant analytic framework of **univariate activation**. This framework was used in a manner directly analogous to previous process-based behavioral contrasts — only now these contrasts indicated the amount by which the key control-process manipulation increased aggregate activity within a brain region of interest. Early neuroimaging studies demonstrated that lateral prefrontal cortex (LPFC), dorsomedial frontal cortex (DMFC), and other nodes within frontoparietal and cingulo-opercular networks (FPN, CON), reliably "activate" during situations of controlled processing [35], a result which has been replicated many times over (e.g., [36-39]).

But the classical approach and its focus on measuring control processes — while serving a necessary, anchoring role — is incomplete. Many cognitive control theories are predominantly specified, not in terms of processes, but in terms of representations ([9]; e.g., abstract rules [3,40], high-dimensional [41] task conjunctions [42,43], compositional sub-tasks [44,45], attentional templates [46], hierarchical task schemas [47-49]). In theory, cognitive control representations, encoded and maintained by control networks (FPN, CON), contain the requisite information for performing cognitive control tasks within an appropriately organized form. But investigating the form of these representations, though it is a central goal, is often challenging to pursue with only process-level measures. Possibly, even, the kinds of questions that would be most clarifying for understanding representational format, are those which process-level measures least naturally support. For example, what makes two control-demanding states similar, versus distinct?

Representational approaches explicitly model control representations

Distinguishing among multiple, similarly control-demanding states is exactly the sort of problem that MVPA methods can make more tractable. These methods can be decomposed (non-comprehensively) into two variants: classification-based decoding, which we refer to here as "classification", and RSA (Box 1; see also encoding methods [50]). There has been a growing body of work using classification within the domain of cognitive control (e.g., for a review [51]), including contributions from our own group ([45,52,53]). But, in contrast to other domains of cognitive neuroscience, (e.g., object recognition or episodic memory; [54-56]), relatively few cognitive control studies have adopted the RSA approach.

We speculate that classification has been used more frequently out of tradition. Classification naturally aligns with the classical approach to cognitive control, as it is well-suited for binary questions — that is, to classify task conditions along a single dimension

(e.g., congruency; see Box 1). In contrast, in its original formulation, RSA was developed for the **condition-rich** naturalistic experimental designs that are common in visual neuroscience [16,57]. Control-oriented researchers may have therefore considered RSA to be more of an exploratory tool for studying sensory and perceptual processes, rather than higher cognitive ones. Yet this would not be accurate. RSA is a general framework in which the central goal is to adjudicate between several competing representational models. Two features of RSA make this a straightforward task: The first is flexibility, due to operating on the **similarity structure** of activity patterns, or their **geometry**, rather than on activity patterns per se. The second is explicitness in the goal of modeling representations, due to a "forward" direction of inference, from model to brain (see Box 1).

In fact, the balance of these features makes RSA well-suited for cognitive control research. Seemingly at odds with the original goal of leveraging the continuously varying feature-spaces of naturalistic experiments, RSA can also capitalize on the strengths of factorial designs. These designs are, of course, a bread-and-butter approach used in cognitive control experiments, as they allow confounds to be efficiently orthogonalized and interactions studied. **Full factorial RSA** provides a simple yet powerful framework for accessing these benefits within analysis of neural representations (Box 2). Indeed, many extant datasets may be amenable to full factorial RSA (e.g., [58]).

Similarly, cognitive control research can leverage the general inferential strategy of RSA, model comparison, in a more comprehensive manner than perhaps initially anticipated. In the original formulation of RSA, a set of competing models (hypotheses) are fit to an observed geometry (e.g., of a brain region), and evidence for a hypothesis is obtained if its corresponding model clearly provides a better fit. But once fit, the modeled representations (i.e, "coding strengths") can also be compared in a variety of useful ways: particularly, in their ability to explain behavioral measures (e.g., classical contrasts of control), in their sensitivity to superimposed experimental manipulations [59] (e.g., process-level factors), and in their ability to explain modulations in other modeled representations (e.g., in downstream regions or subsequent timepoints; termed representational connectivity analysis [60], see Table 1). In conjunction, these tools can provide a rigorous means to decompose the black-box of control processes — the gap between process manipulations and controlled behavior — into a more mechanistic path that is mediated by the strength of specific representations.

RSA and cognitive control: a collection of exemplary studies

To demonstrate the advantages of the RSA approach to cognitive control investigation more concretely, we review a number of illustrative studies, focusing on human neuroscience, that touch on long-standing issues within this domain. While the questions differ, many of these studies share a primary goal: to validate a mapping between a theorized control representation and measured neural activity. Results, therefore, primarily provide support for existing theory, rather than advancing or exploring new theoretical ideas. But, with this validating work in place, RSA methodology can now be used to directly refine and challenge existing theoretical models of cognitive control. Relevant techniques are summarized within Table 1.

Task-set representation

A central element of many theoretical models of cognitive control are task set representations (e.g., [2,3]; for a review, see [61]). One expansive cognitive theory proposes that, during goal-driven behavior, perceptual, action, and contextual information are intermediately bound into conjunctive representations, so-called “event files”, that are used to guide action selection [43,62]. Support for this account was recently found using RSA and EEG, in a pair of rule-based action-selection studies [63,64]. The task required a manual response (left, right, up, down button-press) to be determined by applying a spatial translation rule (e.g., horizontal, vertical, diagonal) to a spatial stimulus (a dot in one of the screen’s quadrants). Conjunctive coding was modeled as the components in the EEG geometry that were uniquely but stably evoked by each stimulus, rule, and response combination (i.e., non-linear combinations of these factors). As these constituent factors (i.e., stimulus, rule, response) were orthogonalized in the design, full factorial RSA was used to dissociate conjunctive coding from “pure” coding of each factor, similar to decomposition of main effects from their interaction. Further, in a novel methodological extension, a single-trial RSA method was developed, which enabled within-subject brain–behavior relationships to be tested via hierarchical linear models.

Conjunctive coding emerged relatively early following the stimulus (i.e., prior to response coding), was robustly related to trial-by-trial response time, and — uniquely among the coding schemes — explained a defining behavioral marker of event files (the “partial repetition cost”). A follow-up experiment imposed an additional stop-signal manipulation on the design, to assess which, if any, coding scheme was impacted by inhibitory control processes [64]. Strikingly, only conjunctive coding was suppressed, shortly after the stop signal was presented, selectively on successfully stopped trials, with the degree of suppression uniquely predicting stopping success. Conjunctive coding thus reflects an important locus of action selection, that is both proximal to behavior and a target or intermediary of inhibitory control. A useful direction for future work will be to clarify the neuroanatomical generators of this conjunctive EEG code (e.g., via fMRI, neurophysiology, or perturbative approaches, such as TMS).

Another major focus of control research is to understand how task sets are regulated. Most commonly, this has been studied within multi-tasking and task-switching paradigms, in which competing task sets are activated and frequently switched across trials [65,66]. Foundational to this research has been the behavioral switch cost: across consecutive trials, people typically are slower and more error-prone at switching tasks versus repeating them. In theory, cognitive control processes must overcome this task-set “inertia” [67] and rapidly activate a new task representation when needed (“reconfiguration”) [24]. Task representations within frontoparietal cortices (through interactions with striatum) are thought to mediate these dynamics [2,3,61].

A strong correspondence between this cognitive theory of task-switching and frontoparietal coding dynamics was recently demonstrated via RSA and fMRI [68]. By conducting RSA on consecutive trials within a cued task-switching experiment, this study illustrated that the strength of task representations in these regions closely tracked inertial and reconfiguration phenomena: task representations were stronger (higher trial-to-trial similarity) following

repeat versus switch trials (inertia) and were additionally strengthened across the cue-to-target epoch (reconfiguration). Critically, these dynamics were also aligned to behavior, such that stronger reconfiguration of the task representation was associated with reduced switch costs.

Sometimes, though, an irrelevant task set can be a potent competitor, not because of its recent use, but because of stable differences in automaticity. This type of regulation is studied in stimulus–response conflict paradigms such as Stroop. How the brain actually regulates this interference is a matter of ongoing debate. Foundational models of control propose that key mediators of Stroop interference are LPFC rule and attentional template representations (e.g., [3,69]). Recent evidence, however, suggests that cognitive control may play less of a role than originally thought (e.g., [70]; see also [71,72]; but see [73,74]). This debate is undoubtedly complicated by the fact that the classical Stroop interference effect is a highly multi-determined measure. Here, though, the RSA framework can assist, as the Stroop paradigm is amenable to decomposition by full-factorial RSA (Box 2; [58]). Such a decomposition can provide a rich set of coding variables that can be used, in turn, to decompose classical behavioral measures from Stroop (as in [58]), or tracked as a function of interference-modulating manipulations (e.g., proportion congruency; [75]). For example, it is possible that top-down control processes, mediated by LPFC rule coding, are better captured by behavioral measures other than the classical Stroop interference effect (e.g., Stroop-effect variability [76]), or, become crucial only in certain scenarios (e.g., when interference is likely [77] or when task statistics engender habitual responding [78]). Thus, not only can RSA be used to constrain the set of mechanisms that putatively mediate response conflict, but also to clarify how best to measure them within behavior.

Learning latent task structure

Real-world tasks often possess a latent structure that unfolds in time. Learning this structure can facilitate performance and increase the probability of advantageous outcomes by enabling the individual to predict which actions are appropriate in given contexts. Recent accounts have postulated that latent structures are encoded by medial and orbitofrontal cortex, and through interactions with midbrain circuitry, serve to guide the formation of frontoparietal task representations that orchestrate goal-directed behavior (i.e., planning and action preparation) [79,80].

This area, as with many others in cognitive control, is amenable to formal modeling. One useful method of formalizing hypotheses in this domain has been with **artificial neural networks (ANNs)**, which perform tasks via distributed representations [48,81], learned during training. By linking artificial representations to those observed in the brain, neural data can constrain ANN models of cognitive control. Because each of these representations can be high-dimensional, often the simplest way to assess such a link is through their geometry — that is, via RSA ([16]).

This feature of RSA was effectively used in a recent fMRI study [82], in which both human participants and an **ANN** learned about the latent sequential structure present in a pair of everyday tasks: making coffee or tea. Although some actions could be freely chosen (e.g., whether to add sugar or water first), others had to be made based on previous choices (e.g.,

only serve the drink after both ingredients had been added exactly once). Successfully performing such a task requires organizing and maintaining relevant information from preceding timesteps, which the critical “context” layer encoded within the ANN (Figure 2A, left). The similarity structure that emerged within this layer served as the RSA model (Figure 2A, Context Layer Model), and was fit to (the similarity structures of) neural activity patterns recorded from participants during performance of the same task. This model was selectively associated with representations in medial PFC (MPFC; including mid-cingulate cortex) — above and beyond competing models defined from other behaviorally relevant variables (e.g., Figure 2A, Sequence Model). The clear specificity of this result, both to MPFC and to the context model, suggests that MPFC could provide downstream regions with a flexible representation of progress-to-current-goal.

Follow-up studies could further strengthen this interpretation by examining how the strength of the MPFC context representation covaries with task representations in other areas and with other manipulations: for example, coordination with LPFC and premotor cortex may be critical during situations of high response conflict, or with hippocampus during learning. Another clarifying direction could be to make the model competition more intense: for example, by testing alternative representations derived from competing network architectures (e.g., [83,84]). Similarly, recent work that could be useful to integrate here, has used classification and RSA methods to identify unique mechanisms for serial-order control, implicating hippocampal and frontoparietal regions in forming hierarchical and “chunked” representations [85-87].

Domain-general cognitive control

A long-standing interest within the field is to refine the constructs of cognitive control (e.g., [1]). Of particular interest is where to draw the boundaries: which functions should be considered general, commonly engaged by different tasks, and which distinct? A construct validation approach is often used to address this issue: measures from a battery of tasks are collected for each individual, and from the covariance structure of individual differences among the tasks, latent factors are estimated that correspond to the hypothesized constructs. Evidence for the construct is provided if factors are interrelated in predicted ways (e.g., [31,32]). Prior studies of the generality of control functions, however, have yielded mixed results (e.g., [71,73]). But, these studies have almost exclusively used behavioral measures to estimate cross-task factors. Brain activity measures may give additional leverage, as they can provide more proximal, higher-dimensional readouts of neural mechanisms.

Through the use of a novel meta-analytic, cross-task RSA, an important first step towards “neural construct validation” of cognitive control was recently reported [88] (see [51] for a complementary voxel-wise meta-analysis). This work included fMRI datasets spanning three domains: cognitive control, negative affect, and pain. The key region of interest was the anterior midcingulate cortex (aMCC), a region that has seemed functionally ubiquitous [89] and challenging to characterize [90]. The researchers found that aMCC representations converged across disparate pain modalities (thermal, mechanical, visceral) and separately, across different sub-domains of negative affect (visual, social, auditory narrative). In contrast, cognitive control representations not only diverged from these two domains, but

diverged from each other — regardless of whether they were from common or distinct sub-domains ("working memory", "response selection", "response inhibition"). That is, no task-general control representations were detected.

Limitations of this study are important to bear in mind. In particular, site factors (e.g., study location, scanner, research group) confound the manipulations of task domain. Likewise, as their meta-analysis was inherently also a cross-subject RSA (i.e., all pattern similarities were estimated between different subjects), the analysis was not optimized to capture finer-grained representational structure that could be subject-specific (i.e., idiosyncratic [91,92]). Indeed, finer-grained or idiosyncratic response topographies may best define the aMCC signal relevant for cognitive control (as, e.g., suggested by differences in developmental timecourses between domains; c.f., [93,94]). Nevertheless, the potential of cross-task RSA is highly appealing. When tasks form the dimensions of a representational similarity matrix, RSA naturally assesses cross-task convergent and divergent validity at the level of neural (e.g., fMRI BOLD signal) implementation (see also [95]). Indeed, similar to construct validation approaches, several RSA methods have been developed that use unbiased similarity measures in order to explicitly account for measurement error (see Table1; [13,96-98]). More broadly, this cross-task view of representations is also reflected in the recent recommendation to use pattern similarity as a rigorous way to assess replication of task-based fMRI results [99].

Interactions with motivational value

There has been a growing appreciation of the tight relationship between motivation and cognitive control [100,101]. In several recent accounts, deciding whether and how to engage cognitive control is a value-based process, in which the benefits of goal attainment are integrated with the cost of various control strategies [102]. This interaction of motivation and cognitive control is thought to involve dopaminergic signaling and fronto–striatal circuitry [103-107]. One way in which this interaction may manifest is as an increase in the gain (i.e., sharpening) of frontoparietal task rule coding. Indeed, a recent cued task-switching study using MVPA classification, reward incentives (which varied trial-by-trial and were pre-cued) increased the distinctiveness of frontoparietal task-set representations, and these task-set coding changes explained individual differences in performance improvement on incentivized trials [52].

Building on this work, the same question was examined from within the RSA framework, via EEG [108]. Relative to [52], the use of full factorial RSA allowed a richer set of coding variables to be jointly estimated — including task rule (as in [52]), but also two stimulus-feature models (target and distractor) and a response (motor) model. While incentives enhanced target stimulus, response, and rule coding (i.e., all but distractor representations), only rule coding was selectively enhanced during rule updating (i.e., on switch versus repeat trials), and emerged prior to trial onset, consistent with a proactive control mechanism [77]. Completing the link to behavior, individual differences in the strength of this interaction between rule coding and incentive robustly explained the amount by which response latencies decreased as function of incentives, even when controlling for incentive-driven changes in other coding schemes. The results of this elegant study converge with those of

[52], but further suggest that this incentive-driven enhancement in rule coding reflects proactive control [77], precedes other incentive-driven representational changes, and may be dependent on updating, as opposed to other types of control processes (e.g., rule maintenance).

From these foundational results, the research paradigm can be expanded to address additional critical questions regarding the motivational properties of incentivized task contexts, such as the effects of motivational valence (i.e., positive/approach, negative/avoidance), incentive categories (i.e., primary, secondary), subjective preferences, and the timing of incentive cues (i.e., preparatory, target-linked) [109,110]. When used with neuroimaging methods of higher spatial resolution (such as fMRI), RSA offers a means of identifying the neural systems underlying these (potentially massively distributed) EEG codes.

Individual differences

Cognitive control is strongly impacted by individual differences [2,77,111]. Some of this variability may be localized to frontoparietal networks, in which functional differences may emerge from, or manifest as variability in task-relevant representations ([2,112]). Studying individual differences in frontoparietal representations can thus be a powerful test-bed for models of control. Likewise, RSA may provide novel ways of approaching such questions, as illustrated by two recent studies.

A long-standing construct in models of control of visual attention is an attentional template representation, encoded by frontoparietal networks, which contains visual information regarding the current target [46] (e.g., an object for which you are searching). At the heart of these models is a computation of similarity, between template features and objects within the visual field: representations of objects are preferentially enhanced, as a function of similarity to template (see [113] for a review). But similarity is, to some degree, subjective [54]. Subjective perceptual differences — idiosyncrasies in perceptual representations — could therefore plausibly drive individual differences in attentional efficiency.

Though this question may seem slippery, it can be naturally addressed with RSA, as demonstrated in a creative fMRI study [114]. Participants first classified identities of faces linearly morphed between two famous individuals (Figure 2B). In an RSA “fingerprinting” procedure adapted from object perception research [54] — which hinges upon on the “second-order” nature of RSA — individuals’ fusiform face area and lateral PFC were found to encode the stimuli in a format that captured their own idiosyncrasies in perceptual categorizations. Critically, an attentional task was performed next, using these same face-morph stimuli as distractors. Only within right lateral PFC did an individual's neural similarity between a given target and distractor — measured during the initial categorization task — predict the degree to which the distractor disrupted (prolonged) their search. This study demonstrates a tight linkage between perceptual categorization, attentional search, and the representational structure in LPFC, while hinting at a privileged role for LPFC in representing categories when relevant for impending decisions. More generally, this study illustrates the utility of the second-order nature of RSA: by abstracting away from brain activity patterns, towards their geometry, investigators can more directly compare

individuals on cognitive dimensions of interest. Beyond identifying idiosyncratic representations, such an approach could be an effective medium for comparing subgroups (e.g., older versus younger adults), circumventing issues of gross anatomical change.

But sometimes the functional topographic organization of cortex is of interest. A recent study conducted by our own group [115], used genetic contributions to individual differences in functional organization to explore task coding within the FPN. Cross-subject RSA was applied to fMRI images of the N-back working memory task, in order to examine the similarity structure of activation patterns between paired individuals, contrasting monozygotic (identical) and dizygotic (fraternal) twin pairs, as well as non-twin siblings and unrelated pairs. Two contrasting models were compared: one of working memory load, the other of stimulus category. The coding of these dimensions was anatomically specific, such that frontoparietal regions showed higher pattern similarity in pairs with greater genetic similarity, but critically, only for the load model. Moreover, these patterns exhibited functionally relevant individual differences: in related (but not unrelated) pairs, stronger common coding of working memory load was associated with better N-back performance. These results provide clear support for the idea that genetic factors are entwined with the development of cognitive control functions [116], and suggest that these factors are expressed in the task-dependent functional organization of frontoparietal networks. Here, the use of RSA methods provided an efficient way to both compare different coding models and identify behaviorally relevant individual differences in the strength of a given coding scheme.

Concluding Remarks

RSA provides a convenient yet rich framework for decomposing control-related neural activity into measures that better correspond to representational components of theories. Of course, there are still many unknowns regarding the limitations of these tools (Box 3), as well as important open theoretical questions (Outstanding Questions). But as we have attempted to illustrate, the RSA framework has high potential for constraining mechanistic theories of cognitive control. We hope that this Review inspires other investigators working in this area to consider whether an RSA approach might be usefully applied to their own research questions.

Acknowledgements

We thank Julie Bugg, Jackson Colvett, Abhishek Dey, Joel Freund, Hannah Maybrier, Emily Streeper, and members of the Cognitive Control and Psychopathology laboratory for providing useful comments that improved an earlier version of this manuscript. We greatly appreciate the constructive feedback and suggestions provided by manuscript reviewers. Funding support for this work was provided to TSB through NIH grant R37 MH066078.

Glossary

Artificial neural network (ANN):
computing systems, loosely based on biological brains (with units analogous to neurons and weights analogous to synaptic connections) that are trained to perform particular tasks via supervised or reinforcement learning algorithms.

Cognitive control:
the coordination and regulation of thoughts and actions in accordance with internally maintained behavioral goals.

Condition-rich RSA:
an experiment containing diverse and high-dimensional experimental stimuli, such as naturalistic images, that permit description by several different, often continuously varying, feature spaces (e.g., Gabor filters, semantic dimensions). A condition-rich RSA leverages this stimulus diversity to disentangle models built on competing feature spaces.

Full factorial RSA:
an RSA approach that uses a combination of a full factorial design (i.e., with multiple fully crossed factors) and multiple regression to decompose various coding schemes and potentially their interactions.

Multivariate Pattern Analysis (MVPA):
A loose category of data analyses that are sensitive to spatially distributed (e.g., across-voxel, across-electrode, or across-neuron) patterns of brain activity. These include but are not limited to classification-based decoding and RSA.

Process:
an account of the function of a neural or cognitive unit (e.g., neuron, area, model component) in terms of the unit's outcome or impact on other systems, or more general contexts in which it is engaged. Analogous to a computing function, which is not described by a set of internal state values (representations), but an abstract operation that "acts on" other values.

Representational geometry:
a term equivalent to similarity structure, but that more strongly connotes the geometric or graphical interpretation (of points as activity patterns and similarity as inter-point distances in high-dimensional activity space), and that emphasizes connections with neural population coding frameworks.

Representation:
an account of the function of a neural or cognitive unit (e.g., neuron, area, model component) in terms of the content and format of information that the unit stably encodes. In the black-box computing function analogy, a representation would be the (hidden) internal values (or states) of the function.

Similarity structure:
Given brain activity patterns evoked by a set of task conditions, the set of all pairwise similarities among the conditions. See also **representational geometry**.

Task set:
The "instructions" for a task — containing information about stimuli, responses, and rules (e.g., stimulus–response mappings) — but represented within a proceduralized, actionable format.

Univariate analysis:

Neuroimaging analysis technique developed to detect spatially uniform (“overall” or mean-level) changes in brain activity resulting from experimental manipulations.

References

1. Diamond A (2013) Executive Functions. Annu. Rev. Psychol 64, 135–168 [PubMed: 23020641]
2. Kane MJ and Engle RW (2002) The role of prefrontal cortex in working-memory capacity, executive attention, and general fluid intelligence: An individual-differences perspective. Psychon. Bull. Rev 9, 637–671 [PubMed: 12613671]
3. Miller EK and Cohen JD (2001) An Integrative Theory of Prefrontal Cortex Function. Annu. Rev. Neurosci 24, 167–202 [PubMed: 11283309]
4. Barch DM and Ceaser A (2012) Cognition in schizophrenia: core psychological and neural mechanisms. Trends Cogn. Sci 16, 27–34 [PubMed: 22169777]
5. Smucny J et al. (2019) Cross-diagnostic analysis of cognitive control in mental illness: Insights from the CNTRACS consortium. Schizophr. Res 208, 377–383 [PubMed: 30704863]
6. Cole MW et al. (2014) The Frontoparietal Control System: A Central Role in Mental Health. The Neuroscientist 20, 652–664 [PubMed: 24622818]
7. deCharms RC and Zador A (2000) Neural Representation and the Cortical Code. Annu. Rev. Neurosci 23, 613–647 [PubMed: 10845077]
8. Dennett DC (1989) The Intentional Stance, MIT Press.
9. Wood JN and Grafman J (2003) Human prefrontal cortex: Processing and representational perspectives. Nat. Rev. Neurosci 4, 139–147 [PubMed: 12563285]
10. Cohen JD (2017) Cognitive Control. In The Wiley Handbook of Cognitive Control pp. 1–28, John Wiley & Sons, Ltd
11. Norman KA et al. (2006) Beyond mind-reading: multi-voxel pattern analysis of fMRI data. Trends Cogn. Sci 10, 424–430 [PubMed: 16899397]
12. Haxby JV et al. (2014) Decoding Neural Representational Spaces Using Multivariate Pattern Analysis. Annu. Rev. Neurosci 37, 435–456 [PubMed: 25002277]
13. Diedrichsen J and Kriegeskorte N (2017) Representational models: A common framework for understanding encoding, pattern-component, and representational-similarity analysis. PLOS Comput. Biol 13, e1005508 [PubMed: 28437426]
14. Cunningham JP and Yu BM (2014) Dimensionality reduction for large-scale neural recordings. Nat. Neurosci 17, 1500–1509 [PubMed: 25151264]
15. Edelman S et al. (1998) Toward direct visualization of the internal shape representation space by fMRI. Psychobiology 26, 309–231
16. Kriegeskorte N (2008) Representational similarity analysis – connecting the branches of systems neuroscience. Front. Syst. Neurosci 2, 1–28 [PubMed: 18958245]
17. Kriegeskorte N and Kievit RA (2013) Representational geometry: Integrating cognition, computation, and the brain. Trends Cogn. Sci 17, 401–412 [PubMed: 23876494]
18. Stroop JR (1935) Studies of interference in serial verbal reactions. J. Exp. Psychol 18, 643–662
19. MacLeod CM (1991) Half a century of research on the Stroop effect: An integrative review. Psychol. Bull 109, 163–203 [PubMed: 2034749]
20. Carter CS (1998) Anterior Cingulate Cortex, Error Detection, and the Online Monitoring of Performance. Science 280, 747–749 [PubMed: 9563953]
21. Rosvold HE et al. (1956) A continuous performance test of brain damage. J. Consult. Psychol 20, 343–350 [PubMed: 13367264]
22. Sudevan P and Taylor DA (1987) The cuing and priming of cognitive operations. J. Exp. Psychol. Hum. Percept. Perform 13, 89–103 [PubMed: 2951490]
23. Meiran N (1996) Reconfiguration of processing mode prior to task performance. J. Exp. Psychol. Learn. Mem. Cogn 22, 1423–1442

24. Rogers RD and Monsell S (1995) Costs of a predictible switch between simple cognitive tasks. J. Exp. Psychol. Gen 124, 207–231
25. Jonides J and Nee DE (2006) Brain mechanisms of proactive interference in working memory. Neuroscience 139, 181–193 [PubMed: 16337090]
26. Monsell S and Driver J, eds. (2000) Control of cognitive processes: attention and performance XVIII, MIT Press.
27. Sternberg S (1966) High-Speed Scanning in Human Memory. Science 153, 652–654 [PubMed: 5939936]
28. Gevins A and Cutillo B (1993) Spatiotemporal dynamics of component processes in human working memory. Electroencephalogr. Clin. Neurophysiol. 87, 128–143 [PubMed: 7691540]
29. Posner MI and Snyder CRR (1975) Attention and Cognitive Control. In *In* Solso R (Ed.) Information processing and cognition: The Loyola Symposium pp. 669–682
30. Schneider W and Shiffrin RM (1977) Controlled and automatic human information processing: I Detection, search, and attention. Psychol. Rev. 84, 1–66
31. Engle RW (2002) Working Memory Capacity as Executive Attention. Curr. Dir. Psychol. Sci. 11, 19–23
32. Miyake A et al. (2000) The Unity and Diversity of Executive Functions and Their Contributions to Complex “Frontal Lobe” Tasks: A Latent Variable Analysis. Cognit. Psychol 41, 49–100 [PubMed: 10945922]
33. Shiffrin RM and Schneider W (1977) Controlled and automatic human information processing: II Perceptual learning, automatic attending and a general theory. Psychol. Rev 84, 127–190
34. Donders FC (1868) Over de snelheid van psychische processen. Onderz. Gedaan Het Physiol. Lab. Utrechtsche Hoogeschool 1968–1869 2, 92–120
35. Cabeza R and Nyberg L (2000) Imaging cognition II: An empirical review of 275 PET and fMRI studies. J. Cogn. Neurosci 12, 1–47
36. Nee DE et al. (2007) Interference resolution: Insights from a meta-analysis of neuroimaging tasks. Cogn. Affect. Behav. Neurosci 7, 1–17 [PubMed: 17598730]
37. Niendam TA et al. (2012) Meta-analytic evidence for a superordinate cognitive control network subserving diverse executive functions. Cogn. Affect. Behav. Neurosci 12, 241–268 [PubMed: 22282036]
38. Assem M et al. (2020) A Domain-General Cognitive Core Defined in Multimodally Parcellated Human Cortex. Cereb. Cortex 30, 4361–4380 [PubMed: 32244253]
39. Fedorenko E et al. (2013) Broad domain generality in focal regions of frontal and parietal cortex. Proc. Natl. Acad. Sci 110, 16616–16621 [PubMed: 24062451]
40. Cohen JD et al. (1990) On the control of automatic processes: A parallel distributed processing account of the Stroop effect. Psychol. Rev 97, 332–361 [PubMed: 2200075]
41. Badre D et al. (2021) The dimensionality of neural representations for control. Curr. Opin. Behav. Sci 38, 20–28 [PubMed: 32864401]
42. Rigotti M et al. (2013) The importance of mixed selectivity in complex cognitive tasks. Nature 497, 585–590 [PubMed: 23685452]
43. Hommel B (2004) Event files: feature binding in and across perception and action. Trends Cogn. Sci 8, 494–500 [PubMed: 15491903]
44. Yang GR et al. (2019) How to study the neural mechanisms of multiple tasks. Curr. Opin. Behav. Sci 29, 134–143 [PubMed: 32490053]
45. Cole MW et al. (2011) Rapid Transfer of Abstract Rules to Novel Contexts in Human Lateral Prefrontal Cortex. Front. Hum. Neurosci 5,
46. Desimone R and Duncan J (1995) Neural Mechanisms of Selective Visual Attention. Annu. Rev. Neurosci 18, 193–222 [PubMed: 7605061]
47. Miller GA et al. (1960) Plans and the structure of behavior, Henry Holt and Co.
48. Holroyd CB and Verguts T (2021) The Best Laid Plans: Computational Principles of Anterior Cingulate Cortex. Trends Cogn. Sci 25, 316–329 [PubMed: 33593641]
49. Badre D and Nee DE (2018) Frontal Cortex and the Hierarchical Control of Behavior. Trends Cogn. Sci 22, 170–188 [PubMed: 29229206]

50. Naselaris T et al. (2011) Encoding and decoding in fMRI. Neuroimage 56, 400–410 [PubMed: 20691790]
51. Woolgar A et al. (2011) Multi-voxel coding of stimuli, rules, and responses in human frontoparietal cortex. NeuroImage 56, 744–752 [PubMed: 20406690]
52. Etzel JA et al. (2016) Reward Motivation Enhances Task Coding in Frontoparietal Cortex. Cereb. Cortex 26, 1647–1659 [PubMed: 25601237]
53. Cole MW et al. (2016) The Behavioral Relevance of Task Information in Human Prefrontal Cortex. Cereb. Cortex 26, 2497–2505 [PubMed: 25870233]
54. Charest I et al. (2014) Unique semantic space in the brain of each beholder predicts perceived similarity. Proc. Natl. Acad. Sci 111, 14565–14570 [PubMed: 25246586]
55. Xue G et al. (2010) Greater Neural Pattern Similarity Across Repetitions Is Associated with Better Memory. Science 330, 97–101 [PubMed: 20829453]
56. Schapiro AC et al. (2012) Shaping of Object Representations in the Human Medial Temporal Lobe Based on Temporal Regularities. Curr. Biol 22, 1622–1627 [PubMed: 22885059]
57. Nili H et al. (2014) A Toolbox for Representational Similarity Analysis. PLoS Comput. Biol 10, e1003553 [PubMed: 24743308]
58. Freund MC et al. (2020) A representational similarity analysis of cognitive control during color-word Stroop. bioRxiv DOI: 10.1101/2020.11.22.392704
59. Popov V et al. (2018) Practices and pitfalls in inferring neural representations. NeuroImage 174, 340–351 [PubMed: 29578030]
60. Coutanche MN et al. (2020) Representational Connectivity Analysis: Identifying Networks of Shared Changes in Representational Strength through Jackknife Resampling, Neuroscience.
61. Sakai K (2008) Task Set and Prefrontal Cortex. Annu. Rev. Neurosci 31, 219–245 [PubMed: 18558854]
62. Hommel B et al. (2001) The Theory of Event Coding (TEC): A framework for perception and action planning. Behav. Brain Sci 24, 849–878 [PubMed: 12239891]
63. Kikumoto A and Mayr U (2020) Conjunctive representations that integrate stimuli, responses, and rules are critical for action selection. Proc. Natl. Acad. Sci 117, 10603–10608 [PubMed: 32341161]
64. Kikumoto A and Mayr U (2020) The Role of Conjunctive Representations in Regulating Actions. bioRxiv DOI: 10/ghv7kp
65. Monsell S (2017) Task set regulation. In The Wiley handbook of cognitive control pp. 29–49, Wiley Blackwell
66. Kiesel A et al. (2010) Control and interference in task switching—A review. Psychol. Bull 136, 849–874 [PubMed: 20804238]
67. Allport DA et al. (1994) Shifting intentional set: Exploring the dynamic control of tasks. In Attention and performance 15: Conscious and nonconscious information processing pp. 421–452, The MIT Press
68. Qiao L et al. (2017) Dynamic Trial-by-Trial Re-Coding of Task-Set Representations in Frontoparietal Cortex Mediates Behavioral Flexibility. J. Neurosci. 37, 0935–17
69. Cohen JD et al. (2000) Anterior cingulate and prefrontal cortex: who's in control? Nat. Neurosci 3, 421–423 [PubMed: 10769376]
70. Moss ME et al. (2020) Does conflict resolution rely on working memory? J. Exp. Psychol. Learn. Mem. Cogn 46, 2410–2426 [PubMed: 31916832]
71. Rey-Mermet A et al. (2019) Is executive control related to working memory capacity and fluid intelligence? J. Exp. Psychol. Gen 148, 1335–1372 [PubMed: 30958017]
72. Algom D and Chajut E (2019) Reclaiming the Stroop Effect Back From Control to Input-Driven Attention and Perception. Front. Psychol 10, 1683 [PubMed: 31428008]
73. Engle RW (2018) Working Memory and Executive Attention: A Revisit. Perspect. Psychol. Sci 13, 190–193 [PubMed: 29592654]
74. Hood AVB and Hutchison KA (2021) Providing goal reminders eliminates the relationship between working memory capacity and Stroop errors. Atten. Percept. Psychophys 83, 85–96 [PubMed: 33165733]

75. Braem S et al. (2019) Measuring Adaptive Control in Conflict Tasks. Trends Cogn. Sci 23, 769–783 [PubMed: 31331794]
76. Williams DR et al. (2019) Putting the Individual into Reliability: Bayesian Testing of Homogeneous Within-Person Variance in Hierarchical Models, PsyArXiv.
77. Braver TS (2012) The variable nature of cognitive control: A dual mechanisms framework. Trends Cogn. Sci 16, 106–113 [PubMed: 22245618]
78. deBettencourt MT et al. (2019) Real-time triggering reveals concurrent lapses of attention and working memory. Nat. Hum. Behav 3, 808–816 [PubMed: 31110335]
79. Niv Y (2019) Learning task-state representations. Nat. Neurosci 22, 1544–1553 [PubMed: 31551597]
80. Holroyd CB and Yeung N (2012) Motivation of extended behaviors by anterior cingulate cortex. Trends Cogn. Sci 16, 122–128 [PubMed: 22226543]
81. McClelland JL et al. (1986) Parallel Distributed Processing.
82. Holroyd CB et al. (2018) Human midcingulate cortex encodes distributed representations of task progress. Proc. Natl. Acad. Sci 115, 6398–6403 [PubMed: 29866834]
83. Herd SA et al. (2013) Strategic Cognitive Sequencing: A Computational Cognitive Neuroscience Approach. Comput. Intell. Neurosci 2013, e149329
84. Kriete T et al. (2013) Indirection and symbol-like processing in the prefrontal cortex and basal ganglia. Proc. Natl. Acad. Sci 110, 16390–16395 [PubMed: 24062434]
85. Kikumoto A and Mayr U (2018) Decoding hierarchical control of sequential behavior in oscillatory EEG activity. eLife 7, e38550 [PubMed: 30426926]
86. Wen T et al. (2020) Hierarchical Representation of Multistep Tasks in Multiple-Demand and Default Mode Networks. J. Neurosci 40, 7724–7738 [PubMed: 32868460]
87. DuBrow S and Davachi L (2014) Temporal Memory Is Shaped by Encoding Stability and Intervening Item Reactivation. J. Neurosci 34, 13998–14005 [PubMed: 25319696]
88. Kragel PA et al. (2018) Generalizable representations of pain, cognitive control, and negative emotion in medial frontal cortex. Nat. Neurosci. 21, 283 [PubMed: 29292378]
89. Poldrack RA (2011) Inferring Mental States from Neuroimaging Data: From Reverse Inference to Large-Scale Decoding. Neuron 72, 692–697 [PubMed: 22153367]
90. Ebitz RB and Hayden BY (2016) Dorsal anterior cingulate: a Rorschach test for cognitive neuroscience. Nat. Neurosci 19, 1278–1279 [PubMed: 27669987]
91. Haxby JV et al. (2020) Hyperalignment: Modeling shared information encoded in idiosyncratic cortical topographies. eLife 9, e56601 [PubMed: 32484439]
92. Ramírez FM et al. (2020) What do across-subject analyses really tell us about neural coding? Neuropsychologia 143, 107489 [PubMed: 32437761]
93. Casey BJ et al. (2008) The adolescent brain. Dev. Rev 28, 62–77 [PubMed: 18688292]
94. Simons SHP and Tibboel D (2006) Pain perception development and maturation. Semin. Fetal. Neonatal Med 11, 227–231 [PubMed: 16621747]
95. Nakai T and Nishimoto S (2020) Quantitative models reveal the organization of diverse cognitive functions in the brain. Nat. Commun 11, 1142 [PubMed: 32123178]
96. Walther A et al. (2016) Reliability of dissimilarity measures for multi-voxel pattern analysis. NeuroImage 137, 188–200 [PubMed: 26707889]
97. Friston KJ et al. (2019) Variational representational similarity analysis. NeuroImage 201, 115986 [PubMed: 31255808]
98. Cai MB et al. (2019) Representational structure or task structure? Bias in neural representational similarity analysis and a Bayesian method for reducing bias. PLOS Comput. Biol 15, e1006299 [PubMed: 31125335]
99. Hong Y-W et al. (2019) False-positive neuroimaging: Undisclosed flexibility in testing spatial hypotheses allows presenting anything as a replicated finding. NeuroImage 195, 384–395 [PubMed: 30946952]
100. Botvinick M and Braver T (2015) Motivation and Cognitive Control: From Behavior to Neural Mechanism. Annu. Rev. Psychol 66, 83–113 [PubMed: 25251491]

101. Pessoa L (2009) How do emotion and motivation direct executive control? Trends Cogn. Sci 13, 160–166 [PubMed: 19285913]
102. Yee DM and Braver TS (2018) Interactions of motivation and cognitive control. Curr. Opin. Behav. Sci 19, 83–90 [PubMed: 30035206]
103. Kouneiher F et al. (2009) Motivation and cognitive control in the human prefrontal cortex. Nat. Neurosci 12, 939–945 [PubMed: 19503087]
104. Leon MI and Shadlen MN (1999) Effect of Expected Reward Magnitude on the Response of Neurons in the Dorsolateral Prefrontal Cortex of the Macaque. Neuron 24, 415–425 [PubMed: 10571234]
105. Locke HS and Braver TS (2008) Motivational influences on cognitive control: Behavior, brain activation, and individual differences. Cogn. Affect. Behav. Neurosci 8, 99–112 [PubMed: 18405050]
106. Pochon JB et al. (2002) The neural system that bridges reward and cognition in humans: An fMRI study. Proc. Natl. Acad. Sci 99, 5669–5674 [PubMed: 11960021]
107. Watanabe M (1996) Reward expectancy in primate prefrental neurons. Nature 382, 629–632 [PubMed: 8757133]
108. Hall-McMaster S et al. (2019) Reward Boosts Neural Coding of Task Rules to Optimize Cognitive Flexibility. J. Neurosci 39, 8549–8561 [PubMed: 31519820]
109. Yee DM et al. (2016) Humans Integrate Monetary and Liquid Incentives to Motivate Cognitive Task Performance. Front. Psychol 6, 2037 [PubMed: 26834668]
110. Chiew KS and Braver TS (2016) Reward favors the prepared: Incentive and task-informative cues interact to enhance attentional control. J. Exp. Psychol. Hum. Percept. Perform 42, 52–66 [PubMed: 26322689]
111. Braver TS et al. (2010) Vive les differences! Individual variation in neural mechanisms of executive control. Curr. Opin. Neurobiol 20, 242–250 [PubMed: 20381337]
112. Duncan J et al. (2012) Task rules, working memory, and fluid intelligence. Psychon. Bull. Rev 19, 864–870 [PubMed: 22806448]
113. Geng JJ and Witkowski P (2019) Template-to-distractor distinctiveness regulates visual search efficiency. Curr. Opin. Psychol 29, 119–125 [PubMed: 30743200]
114. Lee J and Geng JJ (2017) Idiosyncratic Patterns of Representational Similarity in Prefrontal Cortex Predict Attentional Performance. J. Neurosci 37, 1257–1268 [PubMed: 28028199]
115. Etzel JA et al. (2020) Pattern Similarity Analyses of FrontoParietal Task Coding: Individual Variation and Genetic Influences. Cereb. Cortex 30, 3167–3183 [PubMed: 32086524]
116. Friedman NP et al. (2008) Individual differences in executive functions are almost entirely genetic in origin. J. Exp. Psychol. Gen 137, 201–225 [PubMed: 18473654]
117. Jiang J and Egner T (2014) Using neural pattern classifiers to quantify the modularity of conflict-control mechanisms in the human brain. Cereb. Cortex 24, 1793–1805 [PubMed: 23402762]
118. Pereira F et al. (2009) Machine learning classifiers and fMRI: a tutorial overview. NeuroImage 45, S199–S209 [PubMed: 19070668]
119. Hebart MN and Baker CI (2018) Deconstructing multivariate decoding for the study of brain function. NeuroImage 180, 4–18 [PubMed: 28782682]
120. Haynes JD (2015) A Primer on Pattern-Based Approaches to fMRI: Principles, Pitfalls, and Perspectives. Neuron 87, 257–270 [PubMed: 26182413]
121. Shepard RN and Chipman S (1970) Second-order isomorphism of internal representations: Shapes of states. Cognit. Psychol 1, 1–17
122. Hunt LT et al. (2018) Triple dissociation of attention and decision computations across prefrontal cortex. Nat. Neurosci 21, 1471–1481 [PubMed: 30258238]
123. Bobadilla-Suarez S et al. (2020) Measures of Neural Similarity. Comput. Brain Behav 3, 369–383 [PubMed: 33225218]
124. Allefeld C and Haynes J (2014) Searchlight-based multi-voxel pattern analysis of fMRI by cross-validated MANOVA. NeuroImage 89, 345–357 [PubMed: 24296330]
125. Abeles M et al. (1993) Spatiotemporal firing patterns in the frontal cortex of behaving monkeys. J. Neurophysiol 70, 1629–1638 [PubMed: 8283219]

126. Miller EK (1999) The Prefrontal Cortex: Complex Neural Properties for Complex Behavior. Neuron 22, 15–17 [PubMed: 10027284]

127. Stokes MG et al. (2013) Dynamic coding for cognitive control in prefrontal cortex. Neuron 78, 364–375 [PubMed: 23562541]

128. Asaad WF et al. (1998) Neural Activity in the Primate Prefrontal Cortex during Associative Learning. Neuron 21, 1399–1407 [PubMed: 9883732]

129. Rigotti M et al. (2010) Internal Representation of Task Rules by Recurrent Dynamics: The Importance of the Diversity of Neural Responses. Front. Comput. Neurosci 4, 24 [PubMed: 21048899]

130. Bhandari A et al. (2018) Just above Chance: Is It Harder to Decode Information from Prefrontal Cortex Hemodynamic Activity Patterns? J. Cogn. Neurosci 30, 1473–1498 [PubMed: 29877764]

131. Kaanders P et al. (2020) Medial frontal cortex activity predicts information sampling in economic choice. bioRxiv DOI: 10/ghrswp

132. Naselaris T et al. (2021) Extensive sampling for complete models of individual brains. Curr. Opin. Behav. Sci 40, 45–51

133. Gordon EM et al. (2017) Precision Functional Mapping of Individual Human Brains. Neuron 95, 791–807.e7 [PubMed: 28757305]

134. Smith DM et al. (2021) Light through the fog: using precision fMRI data to disentangle the neural substrates of cognitive control. Curr. Opin. Behav. Sci 40, 19–26 [PubMed: 33553511]

135. Miller JA et al. (2021) Overlooked Tertiary Sulci Serve as a Meso-Scale Link between Microstructural and Functional Properties of Human Lateral Prefrontal Cortex. J. Neurosci 41, 2229–2244 [PubMed: 33478989]

136. Glasser MF et al. (2016) A multi-modal parcellation of human cerebral cortex. Nature 536, 171–178 [PubMed: 27437579]

137. Kragel PA et al. (2018) Representation, Pattern Information, and Brain Signatures: From Neurons to Neuroimaging. Neuron 99, 257–273 [PubMed: 30048614]

138. Barron HC et al. (2016) Repetition suppression: a means to index neural representations using BOLD? Philos. Trans. R. Soc. B Biol. Sci 371, 20150355

139. Bhandari A et al. (2019) , Measuring prefrontal representational geometry: fMRI adaptation vs pattern analysis. , in 2019 Conference on Cognitive Computational Neuroscience, Berlin, Germany

140. Liu J et al. (2020) Stable maintenance of multiple representational formats in human visual short-term memory. Proc. Natl. Acad. Sci 117, 32329–32339 [PubMed: 33288707]

141. Chen G et al. (2021) To pool or not to pool: Can we ignore cross-trial variability in FMRI? NeuroImage 225, 117496 [PubMed: 33181352]

142. Haines N et al. (2020) Learning from the Reliability Paradox: How Theoretically Informed Generative Models Can Advance the Social, Behavioral, and Brain Sciences, PsyArXiv.

143. Palestro JJ et al. (2018) A tutorial on joint models of neural and behavioral measures of cognition. J. Math. Psychol 84, 20–48

144. Schaworonkow N et al. (2015) Power-law dynamics in neuronal and behavioral data introduce spurious correlations. Hum. Brain Mapp 36, 2901–2914 [PubMed: 25930148]

145. Diedrichsen J et al. (2020) Comparing representational geometries using whitened unbiased-distance-matrix similarity. ArXiv200702789 Stat at <http://arxiv.org/abs/2007.02789>

146. Munakata Y et al. (2011) A unified framework for inhibitory control. Trends Cogn. Sci. 15, 453–459 [PubMed: 21889391]

147. Basti A et al. (2020) Multi-dimensional connectivity: a conceptual and mathematical review. NeuroImage 221, 117179 [PubMed: 32682988]

Box 1. Strengths of MVPA classification and RSA

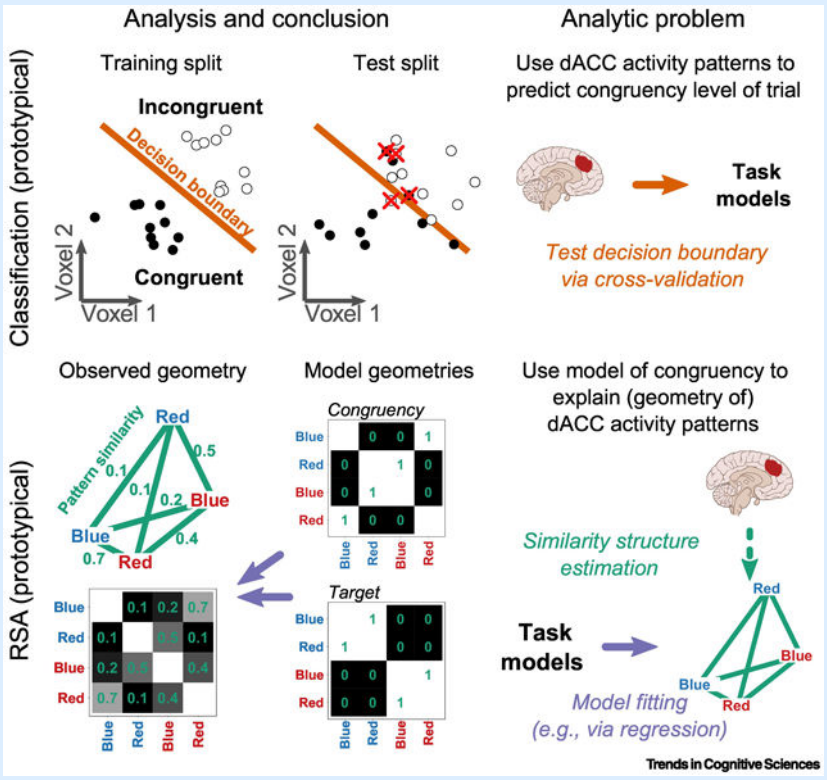
Consider a hypothetical fMRI study of the color-word Stroop task and a researcher interested in the neural coding of congruency information. A one-dimensional research question like this is perhaps most straightforwardly assessed via MVPA classification: can the congruency level of a trial be classified, or "predicted", by patterns of dACC activation (Figure I, Analysis and Conclusion)? If prediction is successful, dACC would be said to encode congruency information (Figure I, test split; see [117] for an actual example of classification in Stroop; for more general discussion and introductions, see [11,118-120]).

But it is often of interest to compare encoding of multiple task variables — for example, does dACC encode congruency selectively or more strongly than task-relevant target information? Here is where RSA becomes advantageous. A prototypical RSA would frame this as a singular problem of model comparison, in which two models (Figure I, congruency and target model geometries) compete to explain a common outcome: the full condition-by-condition similarity matrix of dACC activity patterns (Figure I, observed geometry, lower). The analysis tests whether the congruency model provides a better fit to the observed dACC geometry than the target model.

This example highlights two key distinctions between MVPA classification and RSA. First, RSA operates at a level once-removed from brain activity patterns, whereas classification operates directly on brain activity patterns (Figure I, Analytic Problem). This "second-order" abstraction (a la [121]) is the key basis for the flexibility of RSA. It allows representations from fundamentally different types of spaces to be easily compared [16].

Second, RSA and classification differ in their direction of inference. In RSA, models are tested in their ability to explain (similarity structure of) brain activity patterns (Analytic Problem, RSA, purple arrow), whereas in classification, activity patterns are tested in their ability to predict (or "recover") a hypothesized model (Analytic Problem, Classification, orange arrow; [50]). This "forward" direction of inference of RSA — from models, which are (typically) under experimental control, to brain activity measures, which are not — underlies its explicitness for comparing competing coding models. This is a practical consideration that becomes critical with more complex experiments involving balanced factorial designs (e.g., [122]) that include interactions (e.g., [108]), or when there is a need to control for a diversity of confounding similarity structures (e.g., [63]). Further, this strategy enables one to test whether the geometry is completely explained by the models (i.e., with reference to a noise ceiling; [50,59]).

Finally, many strengths of classification are also accessible via RSA. Similarity measures can be chosen that account for certain properties of the data [123] that popular classifiers handle well, such as structured noise (e.g., [96]). Additionally, RSA models can be fit to individual trials, enabling, for example, using RSA model coefficients to predict theoretically-specified single-trial behavioral indices [63].



Box 1 Figure:

Illustration of prototypical classification and RSA approaches, using a toy example of a four-stimuli color-word Stroop experiment. **Classification, Analysis and Conclusion**, Trial-level activity patterns from a 2-voxel brain region (dACC) are depicted as points in 2-dimensional space. Classification corresponds to fitting a decision boundary (orange) that separates patterns along a single factor (e.g., congruency; hollow or filled points) and is typically assessed via cross-validation (training and test splits). **Classification, Analytic Problem**, Classification directly uses brain activity patterns to predict the task model (congruency). **RSA, Analysis and Conclusion**, Condition-average dACC activity patterns (observed geometry) are depicted as a spatial arrangement (or geometry) of four points (here, the color-word stimuli), with six inter-point distances (green lines). The smaller the distance between patterns, the more similar they are (green numbers). The observed geometry can also be represented as a similarity matrix, as can model geometries. **RSA, Analytic Problem**, In RSA, brain activity patterns are first transformed into geometries, which are then explained by task models (target, congruency).

Box 2: Full factorial RSA

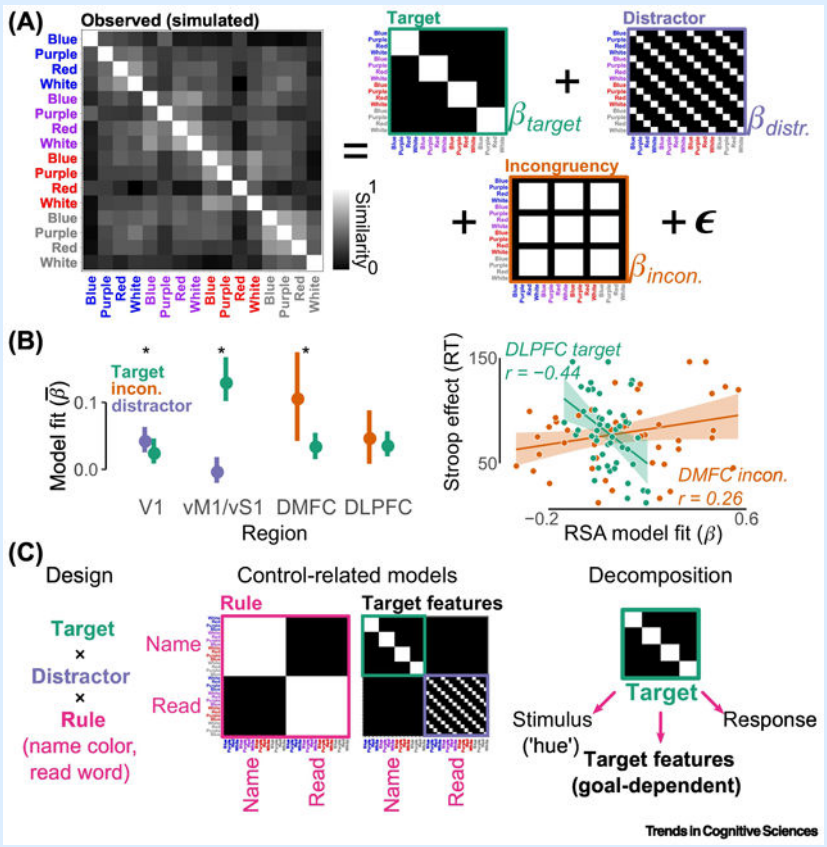
Full factorial RSA offers a convenient framework for removing confounds and studying interactions in neural representations. Yet at first, this approach can seem counterintuitive, particularly for those who have primarily employed factorial designs in univariate contexts. Returning to the example of the Stroop task, we illustrate the potential utility of this approach.

In the color-word Stroop task, the target factor (here, four colors) is crossed with the distractor factor (four words; Figure I, A, left). This design permits at least three coding models to be specified: target, distractor, incongruency (Figure I, A, right). Each model describes an "ideal" coding scheme: for example, the target model would be observed in a region that only encoded the hue (or correct response), regardless of the distractor or congruency. Through multiple regression, these model similarity matrices can be jointly fit to the observed similarity matrix. The resulting β weights reflect the strength with which each factor was uniquely encoded within the observed geometry. Indeed, when this approach was recently, retrospectively applied to fMRI activity from sensory, motor, and frontoparietal control regions, predicted dissociations in coding strength were found at the group level (Figure I, B, left) and in relationships with individual differences in behavior (Figure I, B, right; [58]).

But even this example does not reveal the potential precision of full factorial RSA. For example, the target model conflates a variety of coding schemes: sensory coding (of hue), motor coding (of correct response), and more flexible, attentional template coding (Figure I, C, Decomposition). Adding a rule manipulation (Figure I, C, Design, pink) enables these coding schemes to be unconfounded, as regions encoding an attentional template would be expected to reconfigure their coding scheme on the basis of the task rule (Figure I, C, Control-Related Models, target features model; e.g., [108]; see discussion in [58] for additional decomposition). Likewise, a general rule-coding scheme (Figure I, C, rule model) can now be specified to identify task-set representations.

Further, full factorial RSA allows interaction hypotheses to be tested. For example, if increasing the frequency of incongruent trials drives subjects to adopt a mode of proactive control [77], a corresponding increase should be seen the strength of rule coding. This interaction could be tested by comparing the strength of rule coding (Figure I, C, rule model) in blocks with different probability of incongruency. More complex types of interaction hypotheses are testable, too (e.g., [63]; see also [124]).

An important limitation of full factorial designs are constraints on experimental time. Because time is limited, each additional manipulation typically reduces the number of trials per condition. To some extent, though, a larger RSA matrix (due to having more conditions) will counteract instability due to fewer trials. Nevertheless, manipulations should be added judiciously, boosting precision only along specific dimensions. We strongly recommend piloting and simulation to guide concrete design choices.



Box 2 Figure:

A decomposition of color-word Stroop via full factorial RSA. **A**, The similarity structure evoked during a Stroop experiment is modeled as a weighted sum of three hypothesized coding schemes (for visibility, white-hued stimuli are displayed in grey). **B**, Predicted dissociations in coding schemes were found when applying this approach to fMRI data [58]. At the group level (left), target coding predominated in ventral somatomotor cortex (vM1/vS1), whereas distractor coding predominated in V1 (error bars indicate 95% CI of between-subject variance; asterisks indicate significant pairwise model comparison). Relative to DLPFC, coding of incongruency predominated in dorsomedial frontal cortex (DMFC, including dACC and pre-SMA). At the individual level (right), subjects with stronger target coding in DLPFC, but weaker congruency coding in DMFC had smaller Stroop effects. **C**, In full factorial RSA, the precision of models can be boosted by adding specific manipulations to better isolate representations relevant to cognitive control.

Box 3. Effective fMRI measurement of cognitive control representations

Cognitive control is complex, and perhaps not coincidentally, the brain regions associated with it, such as DLPFC, are challenging to study. Neurophysiologists have long appreciated the difficulty in characterizing the response profiles of neurons within this region (e.g., [125,126]): neuronal selectivities seem to be dynamic, often changing within single-trials [127], and are highly conjunctive, reflecting a mixture of task attributes [128]. At larger spatial scales, the principles by which lateral PFC is functionally organized have been difficult to establish (e.g., [49]). In fact, a recent, influential model of DLPFC embraces the confounding nature of this region by positing that neural populations are randomly connected to their input layers [42,129]. Random connectivity implies a lack of topographic organization, which would cast doubt on the utility of fMRI pattern analysis methods for identifying localized task representations in DLPFC.

Indeed, recent evidence suggests the signal-to-noise ratio in PFC fMRI activation patterns is relatively low. A meta-analysis of MVPA classification studies concluded that mean classification accuracies were considerably lower in PFC than in posterior sensory cortex [130]. A pair of recent RSA studies also illustrates the potential issue. In the first, macaques were trained on a reward-based decision-making task while single-unit activity was recorded [122]. Full factorial RSA revealed a triple dissociation of coding schemes across PFC regions. Yet this same full factorial RSA design, when adapted for human fMRI, found considerably weaker results in putatively homologous regions [131].

Although the scale of topographic organization (or lack thereof) in PFC is one important open question, there are many other potential limiting factors that also likely impact our ability to measure prefrontal coding effectively. More research on the importance of these factors is needed. For instance, dense sampling approaches (fewer subjects, more data per subject) are critical for maintaining signal in the presence of strong individual variability in functional organization [132]. Yet such an approach is rare in cognitive control research, despite the pronounced sensitivity of the associated frontoparietal brain regions to individual differences [133,134]. Thus, individual-specific areal definitions (e.g., [135,136]) are another important avenue to examine. Conversely, expanding the spatial scale of analysis, from areas to networks, has proven highly effective in other domains [137]. Of course, a loss of anatomical precision necessarily comes with this expansion; but, for many cognitive-level inferences, areal versus network-level specificity may not be critical. Finally, repetition suppression may be an effective means to achieve sub-voxel-level precision in measuring representations [138,139].

Importantly, within each of these alternative approaches, the logic of RSA — explicitly modeling neural similarity structure — is applicable. Thus, for those interested in PFC coding of control, we predict that fluency in RSA will be valuable, regardless of measurement technique.

Outstanding Questions

- In stimulus–response interference paradigms such as Stroop, how does cognitive control state (e.g., proactive/reactive) or other contextual and state factors (e.g., trial history, proportion congruency, amount of practice) impact target, distractor and task-set (rule) representations?
- Can more detailed representational models of dACC “congruency” coding (based on, e.g., model-derived response conflict, performance-monitoring information, or value computations) be used to shed light on function of this region?
- Can RSA be used to model the trial-by-trial dynamics of task structure learning?
- What is the evidence for domain-generality (i.e., cross-task similarity) of cognitive control representations?
- How do various motivational factors (e.g., valence, incentive type, preferences) modulate cognitive control representations? Is it through a common mechanism of sharpened task-set coding?
- To what degree are cognitive control representations idiosyncratic, that is, serving as a “fingerprint” of the individual?
- What are the psychometric properties of RSA with regard to individual differences analyses (e.g., test–retest reliability)? Can neuroimaging-based RSA be optimized to more powerfully address individual difference questions?
- Within fMRI-based RSA, what methods (e.g., involving amount of data per subject, preprocessing decisions, region definitions, denoising procedures, similarity estimation) are most effective for measuring prefrontal coding?

Highlights

- Classical measures of cognitive control often only weakly correspond to the theoretical representations they are commonly used to test.
- Representational similarity analysis (RSA) can help better align measures to theory in this domain.
- The power of RSA comes from its flexibility, yet explicitness, in modeling representational structure.
- Full factorial RSA also enhances inferential precision and enables interactions to be tested.
- Useful strategies for applying RSA to inform cognitive control theory are discussed, and recent studies that exemplify these strategies are reviewed.

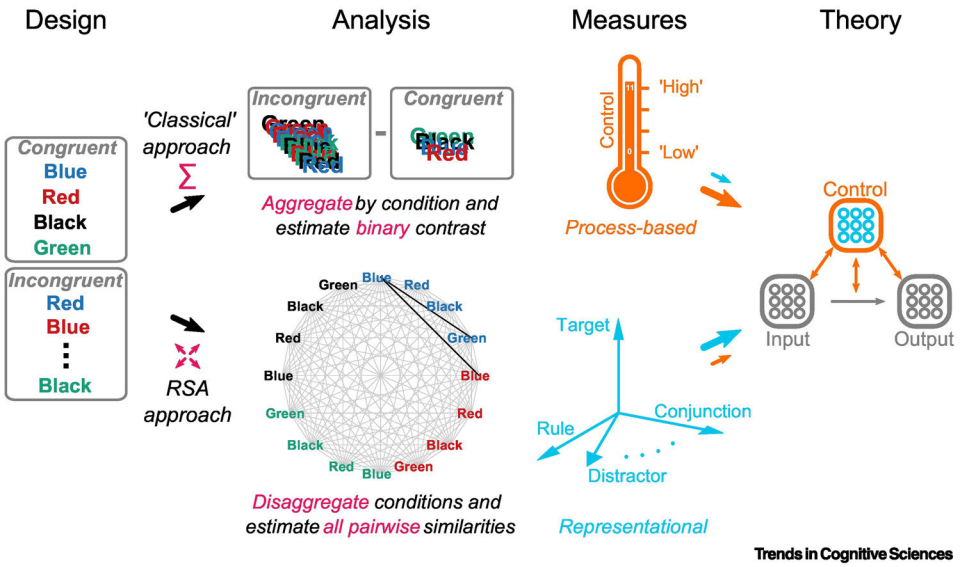


Figure 1, Key Figure.

Schematic of “classical” and RSA-style approaches in cognitive control research. **Design**, A color-word Stroop task with four colors and four words. The participant is instructed to name the stimulus hue rather than read the written word. **Analysis, top**, The classical approach begins by defining abstract factor levels (here, congruent and incongruent) to which conditions (e.g., stimuli) are assigned. Within these levels, the outcome variables of interest (e.g., response time) are summed (Σ) then contrasted. **Measures, top**, These unidimensional measures are typically interpreted in terms of control processes (e.g., slower reaction time on incongruent relative to congruent trials indicates heightened control demands). **Analysis, bottom**, The RSA approach keeps the task conditions disaggregated (↔↕) in order to examine the set of pairwise similarities among measures (e.g., brain activity patterns) from all conditions — that is, their full similarity structure (grey and black lines). This observed similarity structure is then compared to structures predicted from theory. For example, a model of target representations would predict greater similarity between patterns from trials in which the target response was identical (i.e., between stimuli of same hue: e.g., black line connecting blue-hued “BLUE” and “GREEN” stimuli) versus different (e.g., between stimuli of same word in different hues: black line connecting red-hued and blue-hued “BLUE”s). The RSA approach thus provides indices reflecting the strength which multiple different representational schemes were encoded (e.g., the space defined by the light blue basis vectors, which correspond to potentially encoded variables). **Theory**, Typically, classical measures support inference (large orange arrow linking Measure to Theory) regarding control processes (entire CONTROL component of model, orange). Conversely, RSA-based measures can map more directly (large blue arrow linking Measures to Theory) onto theorized control representations (blue nodes within CONTROL component).

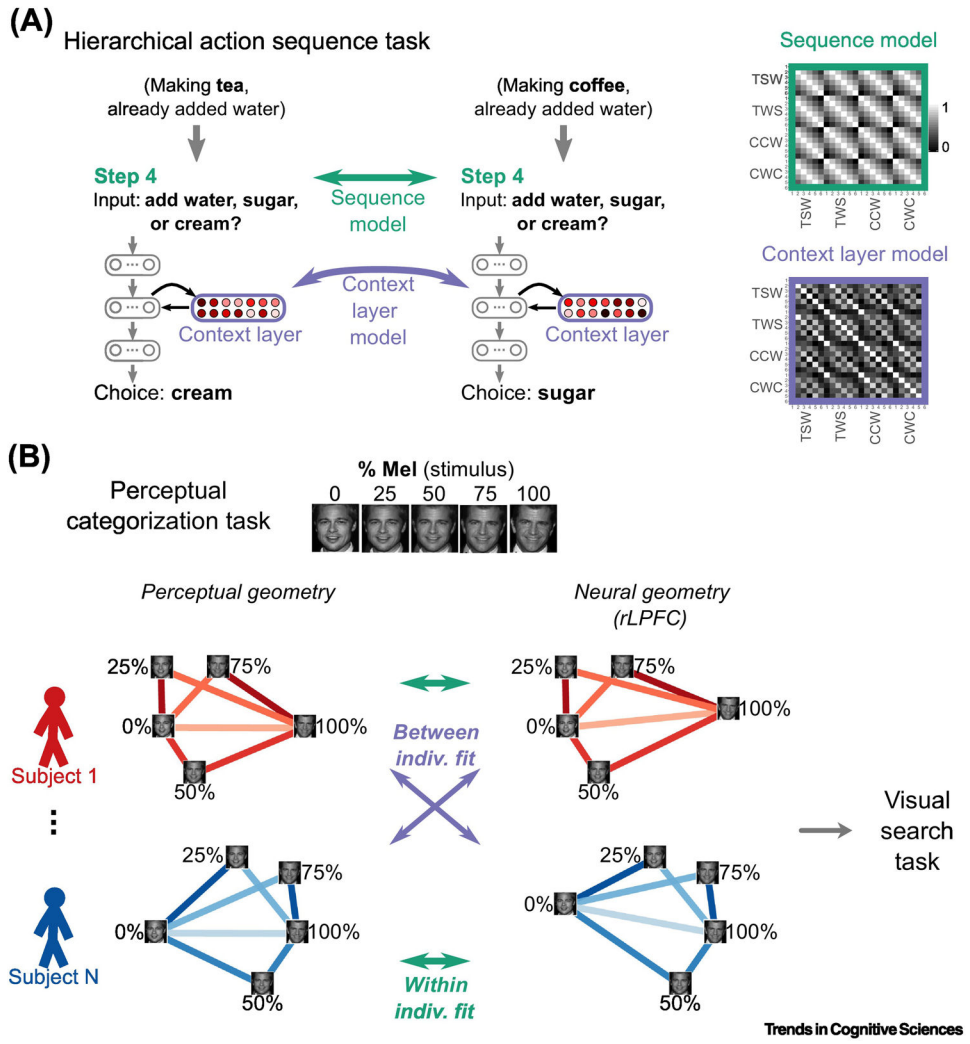


Figure 2.

Diagrams of two reviewed RSA methods. **A**, Mapping internal representations of an artificial neural network (ANN) to brain activity with RSA [82]. An ANN was trained to perform a hierarchical action sequence task, in which the action at one point in the sequence depended on previously chosen actions (e.g., ingredients could only be added once; cream could only be added to coffee). After training, the ANN simulated each step of each sequence (depicted: the fourth step of two different sequences), and the resulting activation patterns (reddish nodes) within the context layer were extracted; the similarity structure of these patterns served as the Context Layer Model (right). A competing model (Sequence Model), which contained only information regarding position-in-sequence (i.e., not previous choices) was built by taking the distance (absolute difference) between each pair of steps (green arrow). **B**, RSA "fingerprinting" [114]. Individuals first performed a famous-face classification task, in which an exemplar face, linearly morphed between two famous faces (e.g., Brad Pitt and Mel Gibson), had to be classified (as either Brad or Mel). Each individual's categorizations were expressed in similarity matrix form (here, depicted as 2-dimensional perceptual geometries) then used as models to explain (green and purple

arrows) neural similarity matrices (neural geometries) from each and every subject. Idiosyncratic brain–behavior relationships were identified in brain regions (i.e., rLPFC) for which the within-subject models (green arrows) were better fit on average than the between-subject models (purple arrows). Neural geometries were then used to predict the patterns of interference within a separate attentional search task that used the same stimulus set (grey arrow).

Table 1.

Summary of various techniques used in RSA.

Technique	Description	Pros (+), Cons (−)	References
ANN-based RSA	*Method of specifying RSA models based on artificial neural network representations.*	+ more direct link between data and formal models + less computationally intensive than estimating unit-to-unit mappings − cannot predict response to novel conditions	• [82], see also [140] • [13,16]
Full-factorial RSA	*Type of experimental approach that uses crossed factors and multiple regression to decompose neural coding.*	+ efficiently boost model specificity + test interaction of representations + handles complex designs − number of trials, experimental time	• [63,108]; see also [122], discussion in [58]; see [124] for related technique
Single-trial RSA	*Analytic procedure for fitting RSA models at single trial level.*	+ test within-subject brain–behavior relationships + supports hierarchical or joint modelling − autocorrelation confounds	• [63,64] • [141-143] • e.g., [144]
Cross-task RSA	*Type of analysis that examines similarity structure within battery of tasks.*	+ assess “neural construct validity” + use to rigorously assess replication	• [88]; see also [95] • [99]
RSA fingerprinting	*Method of assessing presence of stable individual differences in representational structure.*	+ mitigates individual differences due to anatomical factors − requires repeated measures	• [114]; see also [54]
Cross-subject RSA	*Method of assessing task-dependent similarity in response topographies.*	+ useful when individuals can be meaningfully grouped (e.g., by genetic relation) − potential confounds with univariate activity	• [115] see discussion in [92]
Unbiased similarity measures	*Type of similarity measure for which the expected value is not impacted by measurement error.*	+ useful for unbalanced designs − increased variance − each condition must appear in >1 run	• techniques: [13,96,97,124] • unbalanced design: e.g., [75] • increased variance: [145]
Representational “connectivity” analysis	*Type of analysis that examines covariation in coding strength (across region, timepoint). When constrained by RSA models, this covariation is assessed along specific coding variables.*	+ test representational interactions (e.g., b/w PFC and downstream coding + constrained or unconstrained by RSA models − third variables, directionality	• PFC–downstream interactions: e.g., [3,146] • unconstrained: [60,147]; constrained: e.g., see interaction analyses in [108]