Neural Networks 143 (2021) 572–594

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

Internal manipulation of perceptual representations in human flexible cognition: A computational model

Giovanni Granato [a,b], Gianluca Baldassarre [a,*]

[a] Laboratory of Computational Embodied Neuroscience, Institute of Cognitive Sciences and Technologies, National Research Council of Italy, Rome, Italy
[b] School of Computing, Electronics and Mathematics, University of Plymouth, Plymouth, UK

ARTICLE INFO

Article history:
Received 3 September 2020
Received in revised form 30 June 2021
Accepted 9 July 2021
Available online 15 July 2021

Keywords:
Computational model
Goal-directed behaviour
Top-down representation manipulation
Selective attention
Cognitive flexibility

ABSTRACT

Executive functions represent a set of processes in goal-directed cognition that depend on integrated cortical-basal ganglia brain systems and form the basis of flexible human behaviour. Several computational models have been proposed for studying cognitive flexibility as a key executive function and the Wisconsin card sorting test (WCST) that represents an important neuropsychological tool to investigate it. These models clarify important aspects that underlie cognitive flexibility, particularly decision-making, motor response, and feedback-dependent learning processes. However, several studies suggest that the categorisation processes involved in the solution of the WCST include an additional computational stage of category representation that supports the other processes. Surprisingly, all models of the WCST ignore this fundamental stage and they assume that decision making directly triggers actions. Thus, we propose a novel hypothesis where the key mechanisms of cognitive flexibility and goal-directed behaviour rely on the acquisition of suitable representations of percepts and their top-down internal manipulation. Moreover, we propose a neuro-inspired computational model to operationalise this hypothesis. The capacity of the model to support cognitive flexibility was validated by systematically reproducing and interpreting the behaviour exhibited in the WCST by young and old healthy adults, and by frontal and Parkinson patients. The results corroborate and further articulate the hypothesis that the internal manipulation of representations is a core process in goal-directed flexible cognition.

1. Introduction

Executive functions are a set of high-order cognitive processes that allow the expression of flexible behaviour. Diamond (2013) described executive functions as "a family of top-down mental processes needed when you have to concentrate and pay attention, when going on automatic or relying on instinct or intuition would be ill-advised, insufficient, or impossible". This definition largely overlaps with the concept of *goal-directed behaviour* (GDB) and the underlying processes involved (Balleine & Dickinson, 1998; Berman et al., 1995; Daw, Gershman, Seymour, Dayan, & Dolan, 2011; Mannella, Gurney, & Baldassarre, 2013). These processes allow the flexible generation of effective associations between perceptions and actions based on goals and the anticipation of the possible effects that actions might have in the environment before performing them. Goal-directed processes are different from those that underlie *habitual behaviour* (Gläscher, Daw, Dayan, & O'Doherty, 2010; Yin & Knowlton, 2006), which are based on more direct and automatic associations between stimuli and responses.

In the present study, we focused particularly on the goal-directed processes that support *cognitive flexibility*. This is an executive function that allows agents to switch between different behavioural strategies depending on the external and internal conditions. Cognitive flexibility is commonly measured using the *Wisconsin card sorting test* (WCST). This is a neuropsychological test that exists in many forms for humans (Grant & Berg, 1948; Heaton et al., 2000; Milner, 1963; Nelson, 1976) and other species (Brown & Bowman, 2002; Mansouri, Matsumoto, & Tanaka, 2006).

As described in detail in Section 5.3, several computational models have been proposed for studying cognitive flexibility and the WCST (Amos, 2000; Berdia & Metz, 1998; Bishara et al., 2010; Caso & Cooper, 2017, 2020; Dehaene & Changeux, 1991; Kaplan, Şengör, Gürvit, Genç, & Güzeliş, 2006; Levine & Prueitt, 1989; Steinke, Lange and Kopp, 2020; Steinke, Lange, Seer, Hendel and Kopp, 2020b), or for investigating the cognitive processes and neural mechanisms that support executive functions (Ashby, Alfonso-Reese, Waldron, et al., 1998; Gilbert & Shallice, 2002;

* Corresponding author.
E-mail addresses: giovanni.granato@istc.cnr.it (G. Granato), gianluca.baldassarre@istc.cnr.it (G. Baldassarre).

https://doi.org/10.1016/j.neunet.2021.07.013
0893-6080/

Hazy, Frank, & O'Reilly, 2007; Kriete, Noelle, Cohen, & O'Reilly, 2013; Monchi, Taylor, & Dagher, 2000; O'Reilly & Frank, 2006; Rigotti, Ben Dayan Rubin, Wang, & Fusi, 2010; Rougier, Noelle, Braver, Cohen, & O'Reilly, 2005). However, these models are mainly focused on decision-making, response selection, and feedback-dependent learning processes, whereas they overlook the important processes comprising *category learning and category manipulation*. Many experimental and theoretical studies suggest that these processes are at the core of categorisation tasks, of whom the WCST can be considered an instance (e.g., see Seger, 2008; Seger & Miller, 2010; Shankar & Kayser, 2017). Surprisingly, none of the previously proposed models of the WCST have investigated the learning and manipulation of the category representation processes involved in the test despite the fact that performing the test requires that the participants flexibly *switch between alternative category representations*.

In the present study, we addressed this limitation by proposing two novel contributions to offer a new perspective regarding the processes that support flexible cognition and the performance of the WCST. The first contribution is the proposed hypothesis described in detail in Section 3.1, where the *processes related to the top-down internal manipulation of representations and their guidance of goal-directed embodied interactions with the environment play key roles in flexible cognition*. In this study, 'representation' refers to a neural pattern built by a sensory process that extracts relevant abstract features from percepts. Moreover, 'representation internal manipulation' refers to the top-down internal selection and modification of representations in order to extract further information from them that is relevant to the current goals.

This hypothesis builds on two previous contributions from the empirical literature. The first contribution based on studies of goal-directed selective attention processes (Corbetta & Shulman, 2002; Gottlieb, 2007; Miller & Cohen, 2001) shows that goals can drive the top-down selection of relevant parts of sensory input patterns. The second contribution based on investigations of covert attention and imagination (Gazzaley et al., 2008; Gazzaley & Nobre, 2012; Kosslyn, 1999; Mechelli, Price, Friston, & Ishai, 2004) shows how the top-down activation of visual cortices can support the visual working memory and imagination by encoding stimuli that are not currently perceived. The hypothesis is also linked to theories of *embodied cognition* that treat cognitive processes and internal representations as strongly grounded on the interactions of agents with the environment (Barsalou, 2008; Caligiore, Borghi, Parisi, & Baldassarre, 2010; Clark, 1997). Finally, the hypothesis is also driven by recent progress in machine learning related to the development of *generative neural networks* that are capable of producing new 'plausible' images through their internal computations (Goodfellow et al., 2014; Hinton, 2012; Hinton, Osindero, & Teh, 2006; Kingma & Welling, 2013; Le Roux & Bengio, 2008). We envisage the existence of parallelisms between these generative processes and the interactions between the frontal and visual cortices of the brain considered above.

In summary, in our proposed hypothesis, flexible goal-directed cognition and behaviour rely on the capacity of brain to: (a) select categories and use them to internally re-code perceptual stimuli; and (b) use the resulting representations to select and guide the execution of suitable actions that are directed to accomplishing the desired goals.

Our second contribution is the proposal of a new computational model that operationalises the proposed hypothesis. The model includes some processes that are conceptually analogous to those used in previous models, such as a mechanism for selecting the behavioural rule to follow and an error processing mechanism to correct this selection. However, our model also includes a fundamental novel feature derived from our hypothesis: a system of processes that can manipulate internal representations for solving tasks such as the WCST. This system relies on the following three key components: (a) an *executive working memory* for storing possible categorisation rules that the system can use to deferentially interpret perceptual stimuli; (b) a *visual working memory* for representing perceptual stimuli at different levels of abstraction; and (c) an *internal manipulator* that selects the contents of the visual working memory based on the categorisation rules stored in the executive working memory. The synergistic interaction between these components and other auxiliary sensorimotor elements allow the model to flexibly interact with the environment to perform the WCST. Thus, the model can perform the test by actively switching between internal representations of stimuli rather than by simply selecting abstract rules or actions, as found in previous models.

We validated the model by reproducing and accounting for the data obtained from the following different human groups tested with the WCST: early adults, old adults, frontal patients, and Parkinson's patients. Moreover, we performed various 'lesions' of the model by manipulating its key parameters, which allowed us to further clarify the contributions of the cognitive processes reproduced in the model to the production of the behavioural errors typically manifested by the different target groups.

The performance of the WCST does not strictly require the new mechanism for the internal manipulation of representations proposed in this study, as demonstrated by the fact that several other models can perform this task based on other solutions. Nevertheless, it was important to check that the proposed model could perform and account for the WCST at the level of other state-of-the-art models of the task. Indeed, our hypothesis states that the internal manipulation of representations is at the core of flexible cognition and the WCST is considered a major tool for testing this function (Section 5.3). However, it should be noted that our three-component hypothesis extends well beyond the WCST and it represents a general theory regarding the key cognitive processes that underlie flexible cognition. In this respect, our research agenda envisages the use of the hypothesis and models derived from it, including the one proposed in the present study, to reproduce other executive functions such as planning (Diamond, 2013).

2. WCST and brain processes involved in its performance

2.1. Materials and procedure

In the WCST (Fig. 1, top), the participants must match a card drawn from a deck (called a 'response card' or 'deck card') with one of four sample cards (called 'stimulus cards' or 'target cards'). Each card contains coloured items with a unique combination of features. These features are differentiated into three categories and each has four attributes: (1) colour: red, green, blue, or yellow; (2) form: stars, triangles, circles, or crosses; and (3) number: one, two, three, or four elements. The participant is requested to move the deck card close to one of the target cards by trying to match them in terms of either their colour, form, or number. A first key challenge in the test is that the participant is not told the correct rule for matching the cards. After each action, an operator provides 'correct' or 'incorrect' feedback based on the current matching rule. The participant must then infer the correct rule based on this feedback. A second key challenge in the test involves probing cognitive flexibility. The correct matching rule changes after a certain number of uninterrupted correct actions and when this occurs, the participant must search and switch to the new rule based only on the information provided in the feedback. Finally, in order to pass the test, the participant must complete a certain number of uninterrupted card sequences, where each involves a different rule.

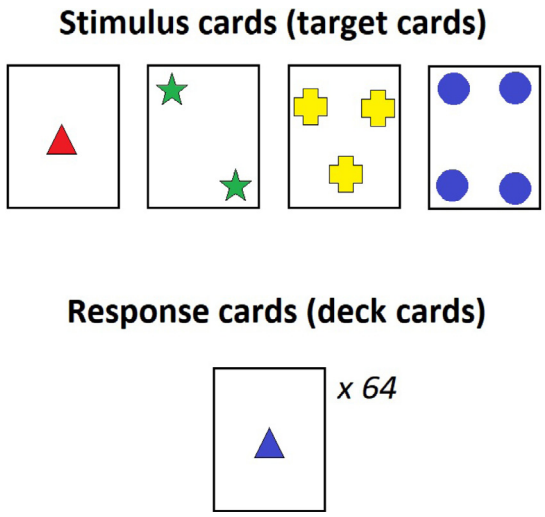

Fig. 1. Schema showing the typical elements in the Wisconsin card sorting test. (For the colours of the card items, the reader is referred to the web version of the article.).

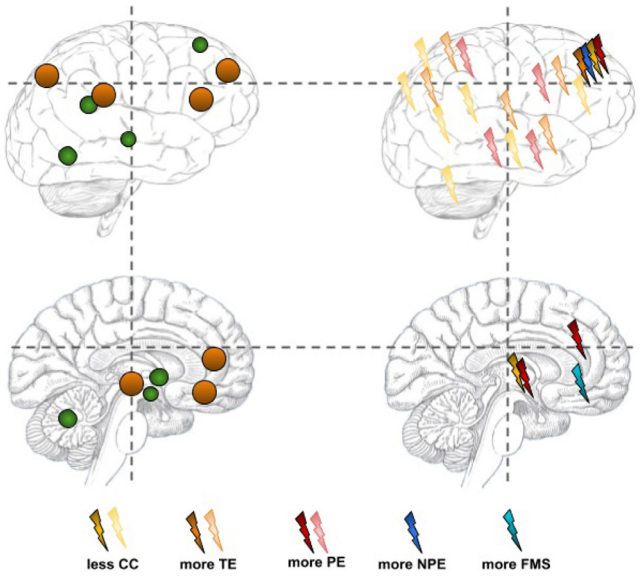

Fig. 2. Left: Highly active brain areas during the performance of the WCST. The colour and size of each circle indicates the number of studies considered that identified a specific activation site (small/green: < 3; large/orange ≥ 3). Right: Sites of lesions that cause specific errors during the WCST. The colour intensities of the bold arrows indicate the specificity of lesions (transparent: distributed lesions; dark: focused lesions). For both graphs, see Table S1 in Supplementary Materials for previous studies that support this overview. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Many versions of the WCST are available with differences in the test procedure or performance score. We used Heaton's version of the test (Heaton et al., 2000) with the following specific features: (a) participants can use up to two decks of 64 cards; (b) completing a category set requires ten correct matches in sequence; (c) after completing a category set, the sorting rule changes but the participant is not told; (d) in order to pass the test, the participant must complete a series involving six different correct matching rules: colour, form, number, colour, form, and number; and (e) if the participant uses both decks without completing the series of categories, the test is considered 'failed'.

2.2. Scoring and types of errors

To score the test, we followed the official documentation for the test (Heaton et al., 2000). In particular, we used the following five principal indices to give a full profile of the participant's performance (see the documentation for thorough explanations of the indices).

- *Completed Categories* (CC): this index ranges from (0, 6) and indicates the number of successfully completed categories to score the global performance.
- *Total Errors* (TE): total incorrect responses, including both perseverative errors and non-perseverative errors (see below), as an index for scoring the level of global deficit.
- *Perseverative Errors* (PEs): cards sorted with the same incorrect rule after a negative feedback error as an index representing perseverative behaviour.
- *Non-Perseverative Errors* (NPEs): errors not included in PEs, where these errors can occur in different situations and they may suggest attentional failure or incorrect inferential reasoning.
- *Failure-to-Maintain Set errors* (FMS): any error that occurs after five consecutive correct matches.

2.3. Neural correlates of behaviour exhibited during the performance of the WCST

Our proposed model has the objective of operationalising the three-component hypothesis. The model reproduces selected features of the architecture and functioning of the brain. These features are at a high level because reproducing specific biological details of the brain anatomy and physiology was beyond the scope of this study.

Previous studies have proposed various partially overlapping interpretations of the neural correlates of the behaviour exhibited by WCST participants.Fig. 2 presents a schematic overview of the brain areas that have a high activation during the performance of the WCST and of lesioned sites linked to specific errors that occur during the test. To build this diagram, we analysed the studies described in the meta-review by Nyhus and Barceló (2009) (see Table S1 in Supplementary Materials for further details).

This analysis indicates that the brain areas that contribute most to the performance of the WCST are the frontoparietal cortices associated with goal-directed perception and attention (Parks & Madden, 2013; Vossel, Geng, & Fink, 2014), sub-cortical structures (particularly basal ganglia) linked with the processing of rewards (Yin, Ostlund, & Balleine, 2008), frontal structures such as the orbital and ventromedial prefrontal cortex (PFC) that support emotional processing, and the anterior cingulate cortex (ACC) associated with error detection (Stuss et al., 2000; Zald & Andreotti, 2010). Several studies (e.g., Goldman-Rakic, 1996; Hoffmann, 2013) indicate that these frontal, parietal, and subcortical systems form an integrated network of systems that underlie cognitive flexibility and other executive cognitive functions.

Fig. 2 also summarises the relationships between errors that occur during the performance of the WCST and different lesion sites. In agreement with the functional interpretations of the active areas observed in healthy participants discussed above, neuropsychological studies have highlighted the presence of (a) a correlation between PEs and a lesion in both the subcortical and medial cortices, and (b) a correlation between most types of errors and a lesion in the superior frontal cortices. In particular, PEs are considered to be indicators of an impairment in flexibility related to the incapacity to change a behavioural rule that has been successful up to a certain point (Dehaene & Changeux,

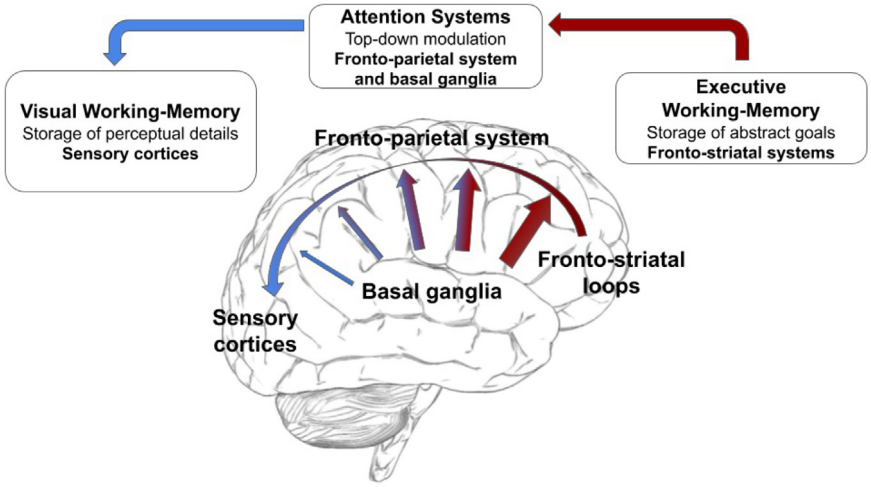

Fig. 3. Schema showing the three-component hypothesis regarding the internal manipulation of representations. The colour gradient (red to blue) indicates a gradual change in the computational functions from those encoding goals and behavioural rules (red: frontostriatal areas) to those encoding percepts (blue: dorsocaudal areas). (For the reference to colours in this figure, the reader is referred to the web version of this article.).

1991; Nelson, 1976). In addition to these classical interpretations focused on PEs, some studies focused on NPEs and FMS errors. In particular, Li (2004) suggested that NPEs are related to attentional or reasoning failures, whereas (Barceló & Knight, 2002) linked them to attention and working memory dysfunctions. Figueroa and Youmans (2013) focused on FMS errors and suggested that they reflect distractibility rather than a cognitive flexibility deficit.

3. Three-component hypothesis and model description

3.1. Three-component hypothesis regarding the flexible internal manipulation of representations

Our hypothesis states that the internal manipulation of representations relies on the interplay among three fundamental brain systems, thereby resulting in the three 'components' of the hypothesis (Fig. 3): (a) a component for storing goals and behavioural rules; (b) a component for manipulating perceptual representations based on the goals and behavioural rules; and (c) a component for extracting perceptual representations and possibly for recalling them based on a bias received from the manipulation component.

These components reproduce the related functions of the brain at an abstract level. The overall top-down modulation of internal representations is a brain mechanism that exerts a bias onto the information flows passing through the cortical pathways (Caligiore, Arbib, Miall, & Baldassarre, 2019; Cisek & Kalaska, 2010). Several studies (e.g., Baldauf & Desimone, 2014; Fuster & Bressler, 2015; Gazzaley & Nobre, 2012; Kosslyn, 1999; Mannella & Baldassarre, 2015; Mechelli et al., 2004) have suggested that this mechanism supports the extraction of external input features, favours the top-down biased selection of relevant information (selective attention), and allows the internal persistent or maintenance of information in the absence of an external input (working memory). By integrating a vast number of previous studies (Corbetta & Shulman, 2002; Fuster & Bressler, 2015; Gottlieb, 2007; Miller & Cohen, 2001; Wolters & Raffone, 2008) and our previous theoretical and modelling studies (Baldassarre, Caligiore and Mannella, 2013; Baldassarre et al., 2013b; Caligiore et al., 2019, 2010; Mannella & Baldassarre, 2015), we now consider how the three components that are operationalised in the computational model might work together to support the performance of the WCST.

The more abstract and amodal *executive working memory* (Braver & Bongiolatti, 2002; Hartley & Speer, 2000) relies on frontostriatal networks. Based on motivational drives, this memory stores the overall goal to pursue (e.g., obtaining positive feedback in the WCST) and the possible sub-goals required to accomplish it (e.g., fulfilling the card matching rules). The goal and sub-goals are encoded as perceptual representations according to the abstraction processes in the perceptual cortical hierarchies.

The *frontoparietal cortical system* is based on perceptual attentional processes (Parks & Madden, 2013; Vossel et al., 2014) and basal-ganglia selection mechanisms (Chelazzi, Perlato, Santandrea, & Della Libera, 2013; Pessoa, 2015; Redgrave, Prescott, & Gurney, 1999; Seger, 2008). This system applies a top-down bias on the lower-level competition processes that occur within the sensory cortices. In particular, this bias is driven by the behavioural rules stored in the executive working memory and it selects alternative contents within the perceptual cortical system (e.g., the colour rather than the shape of the elements in the card).

The *perceptual cortical system* has two functions. First, it transmits sensory information to the higher-level cognitive systems (Gazzaley & Nobre, 2012; Rizzolatti & Matelli, 2003). Second, it is excited by the top-down biases to implement a perceptual working memory (Raffone, Srinivasan, & van Leeuwen, 2014) for storing selected perceptual features related to the accomplishment of the required goals (e.g., specific features of cards based on different possible card sorting rules).

Fig. 4 shows how the three components work in synergy to support the performance of the WCST. The sub-goals comprising the behavioural rules linked to specific categories and stored in the executive working memory stimulate the selection of specific contents in the perceptual working memory. Under this bias, the perceptual working memory generates a representation of the deck card and the target card by emphasising a certain category (e.g., colour). Then the two representations are compared to check whether the two cards match or not. The outcome of this comparison guides the downstream action selection process. Afterwards the system produces the response if the cards match, but if they do not match, another target card is selected to compare with the deck card.

Our hypothesis and model agree with previous research regarding the macrostructure and functional anatomy of the brain. In particular, empirical evidence indicates that the PFC is strongly connected with the parietal cortex to the same extent or even more than with the motor areas (Passingham & Wise, 2012;

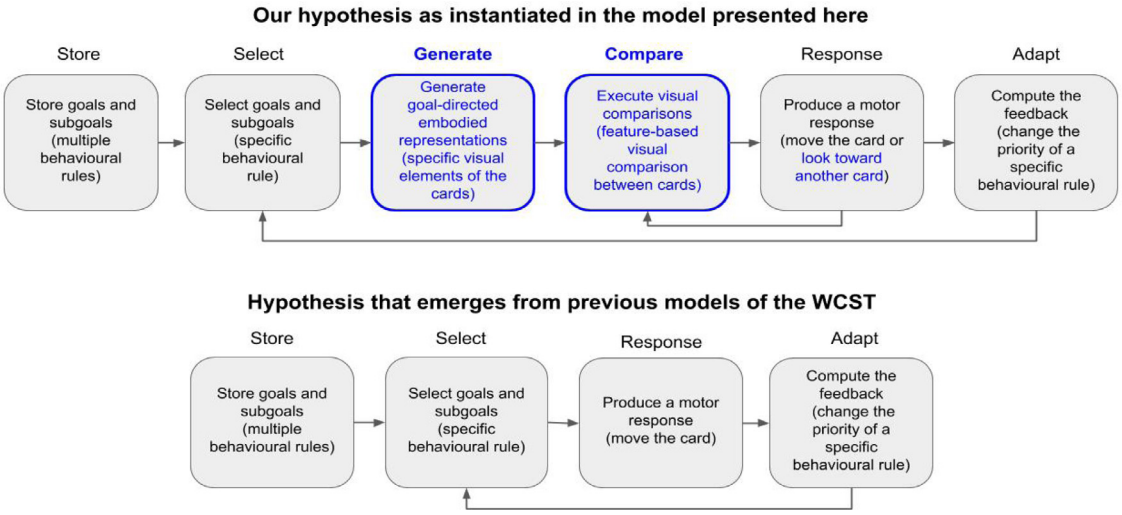

Fig. 4. Top: Key elements of our proposed hypothesis regarding the processes that might underlie flexible cognition and the solution of the WCST. Bottom: Hypothesis based on other models of the WCST.

Rizzolatti & Craighero, 2004). The parietal cortex then exerts strong control on the motor areas. The parietal cortex plays a key role in controlling actions based on representations of the features of objects that are relevant for interacting with them, such as their size and position in space (Jeannerod, Arbib, Rizzolatti, & Sakata, 1995; Thill, Caligiore, Borghi, Ziemke, & Baldassarre, 2013). These representations are considered to encode *affordances* (Gibson, 1979; Norman, 1988), that is, the agent's internal representations of the preconditions necessary for the successful accomplishment of actions (Baldassarre, Lord, Granato, & Santucci, 2019; Fagg & Arbib, 1998; Thill et al., 2013). This idea is at the core of our hypothesis and it contrasts with the view summarised in Fig. 4 based on all previous models of the WCST. According to these models, the high-level selection of the category rule directly biases the selection of the motor responses rather than the lower-level perceptual representations, as stated in our hypothesis.

3.2. Overview of functioning of the components of the model and their biological underpinnings

The architecture of the model was designed based on the organisation of the macro-structural areas of the brain that underlie the functions relevant to our hypothesis: perceptual and category learning, working memory, and the internal selection of representations. The model was abstracted over the anatomical and physiological details of the brain micro-circuits and neurons. This simplification allowed us to realise the first operationalisation of the three-component hypothesis. More biologically plausible implementations of the components might be realised in future work. The architecture also encompasses some auxiliary components that are required for an agent to autonomously form realistic representations of the cards and to interact with the environment. The architecture and components of the model are shown in Fig. 5, and they are now explained in detail.

Visual sensor. This component corresponds to the retina of the eye. The agent actively displaces the sensor so that it focuses on one card at a time (see below). The sensor returns a visual image of the cards. The image is sufficiently large such that a focused card is completely within its scope.

Perceptual component. This component is a layered neural network that performs the bottom-up processing of visual images. The component reflects the hierarchical nature of visual cortices involving many levels of information processing (Baldassarre, Caligiore et al., 2013; Felleman & Van Essen, 1991; Mechelli et al., 2004) ranging from the low-level retinotopic visual processing of features in the striate cortex (V1) to the processing of higher-level image proprieties (shape, colour, etc.) in extra-striate cortices (V2 to V5) and different areas of the inferotemporal and parietal cortex (DeYoe et al., 1996; Konen & Kastner, 2008; Rizzolatti & Matelli, 2003).

Reward/motivation component. This component processes the input to compute the system's internal reward signals and applies a motivational bias to the working memory processes. In the brain, these processes rely on the ventral basal ganglia (Humphries & Prescott, 2010; Mannella et al., 2013), ventromedial PFC, and ventral portion of the ACC (Gläscher et al., 2012; Gläscher, Hampton, & O'Doherty, 2008).

Executive working memory. This component reproduces the functions of the executive working memory, particularly storing the possible sub-goals that correspond to the possible card matching rules. The executive working memory is supported by dorsolateral portions of the PFC, which can store information regarding goals and behavioural strategies through recurrent circuits (Barraclough, Conroy, & Lee, 2004; Braver & Bongiolatti, 2002; Hartley & Speer, 2000), and select them based on lateral inhibitory mechanisms (Aron, 2007). Similar to the working memory of the brain (Brunel & Wang, 2001; Gruber, Dayan, Gutkin, & Solla, 2006), the model executive working memory implements the following key processes (cf. Frank, Loughry, & O'Reilly, 2001; O'Reilly & Frank, 2006): (a) the active maintenance of sub-goals in the absence of perceiving the corresponding stimuli; and (b) releasing (forgetting) this information when it is no longer relevant.

Perceptual manipulator. This component supports two processes. The first process involves decisions regarding behavioural rules that depend on activation of the executive working memory. This process mimics the role of the dorsal ACC and other PFC areas in affective decision making (Bush et al., 2002; Heilbronner & Hayden, 2016; Silvetti, Vassena, Abrahamse, & Verguts, 2018). The second process involves the performance of the actual top-down manipulation of the perceptual contents of the perceptual system. The manipulation is based on a disinhibition mechanism

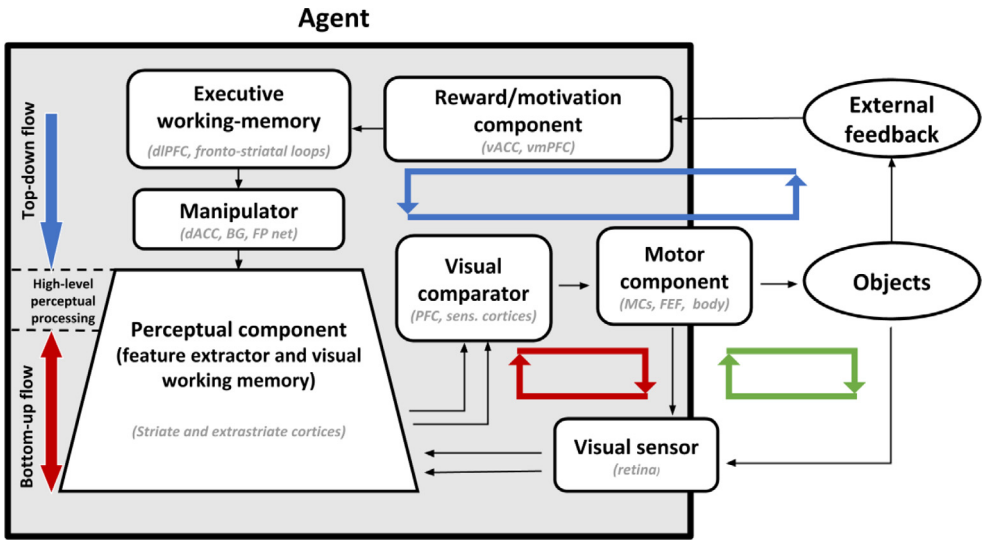

Fig. 5. Schema showing the model components, functions, flows of information between the components, and interaction loops that allow the agent to engage with the environment (red: attentional loop; green: object-displacement loop; blue: feedback-manipulation loop). (For the colours of the interaction loops, the reader is referred to the web version of the article.).

that reproduces the main features of the functioning of basal ganglia (Mannella & Baldassarre, 2015; Redgrave et al., 1999). This mechanism inhibits all internal representations activated by the bottom-up sensory information, except for the one that corresponds to the card matching rule that needs to be followed. Based on this mechanism, only the colour, form, or size features are used to compare the deck and target cards. The manipulator also employs a local selection process to enhance the activation of specific features observed in the stimuli and in agreement with the top-down bias (e.g., to select 'red' if the chosen behavioural rule is 'colour'). This process mimics the top-down modulation effect of the high-level cortices on the lower sensory cortices via the frontoparietal cortical system and basal ganglia (Gazzaley & Nobre, 2012; Parks & Madden, 2013; Vossel et al., 2014; Yin & Knowlton, 2006).

Visual comparator. This component compares the deck card and the foveated target card, and returns their level of similarity ('visual matching'). This component is inspired by findings related to same/not-same tasks (Perani et al., 1999), which have been shown to rely on the interplay between the occipital/temporal cortices and dorsolateral PFC.

Motor components. The first motor component moves the visual sensor to scan the deck card and then the different target cards in order to search for the one that matches the deck card based on the selected matching rule. This component guides the gaze in a top-down manner based on the visual comparator output, as follows. When there is not a visual match, the mechanism triggers a saccade that shifts the fovea to the following target card. When there is a match, the mechanism stops the gaze on the current target card and releases the arm action. When this occurs, the second motor component controls a simulated manipulator that moves the deck card close to the selected target card, as required by the WCST. It is assumed that these attentional scanning and object-moving behaviours are acquired before the solution of the WCST.

Bottom-up perceptual processes and top-down attentional processes. The model implements bottom-up and top-down information flows that correspond to perception and attention processes, respectively (Dijkstra, Zeidman, Ondobaka, Gerven, & Friston, 2017; Intaitė, Noreika, Šoliūnas, & Falter, 2013). Perception involves the bottom-up transmission and progressive abstraction of visual information from the retina to higher cortical levels. Attention and imagination involve a top-down information flow through the frontal areas of the brain that can bias peripheral perceptual areas, and thus they tend to exhibit stronger activation corresponding to relevant external stimuli (attention; Mechelli et al., 2004), or they can even be activated in the absence of them (imagination; Kosslyn, 1999). In the model, the selection of a specific matching rule within the working memory and the consequent disinhibition of a certain attribute representation start a top-down activation flow. This flow leads to the generation of images at the lower levels of the perceptual component that correspond to the selected attribute. In order to perform the task, the model uses these generative processes both with the deck card and the target card under focus. The resulting rule-based representations are then used to compare the two cards at the low perceptual level with respect to the selected colour/shape/size category.

A specific consideration must be made regarding the latter process, as follows. The deck/target card comparison could be performed based on the high-level representations of cards, such as in the last layer of the perceptual component. However, as discussed in Section 3.1, the comparison at the low level is proposed to mimic the functioning of the brain. Moreover, this approach might also have the following computational advantages: (a) the possibility of exploiting the detailed information received from the sensors and selected in a suitable manner by the top-down processes to conduct operations that cannot be performed at a higher level of abstraction (e.g., operations that depend on the detailed shapes of objects; Mechelli et al., 2004; Wolters & Raffone, 2008); and (b) the possibility of using high-level abstract representations to generate lower-level detailed representations based on information gathered 'along the way' while the activation process spreads through the intermediate representation levels, where this more detailed information can then be used by processes that depend on it, such as fine-level comparisons (Barceló, Suwazono, & Knight, 2000; Gazzaley et al., 2008; Mangun, 1995; Woldorff et al., 1997).

Interaction loops. The top-down manipulation mechanism described above is coupled with three interaction loops via the environment (Fig. 5). These loops allow the model to perceive visual stimuli (images) with a substantial level of realism and, most importantly, to use the manipulated representations to support flexible behaviour.

A first 'attentional loop' involves the motion of the visual sensor, which allows the model to observe deck and target cards, thereby affecting the model's internal processing. A second 'object-displacement loop' involves the motor system, which allows the model to displace the deck card, thereby affecting the following visual percepts. A third 'feedback-manipulation loop' allows the model to process feedback to produce an internal reward signal, which is used to update the relevance of the used matching rule stored in the executive working memory. Thus, the third loop affects the operation of the other two loops.

These loops involving circular interactions between the model components and environment capture the essence of the sensorimotor interactions involved in the solution of the WCST. The loops are simple but sufficient to study the proposed top-down perceptual manipulation mechanism. The model does not have a full embodiment, for example it lacks specific actuators with realistic physical dynamics, but it still has some key embodied features. In particular, the actions of the model can affect its sensory input and the active control of this input is part of the strategy used by the model to perform the task. According to some views with which we agree, the fact that the solution to a problem relies on the circular loop where the agent interacts with the environment represents a key element of embodiment (Nolfi & Floreano, 2000).

3.3. Computational details of the model

The key computational features of the model are summarised in Fig. 6. Further details and equations regarding the functioning and learning of the model are presented in the Supplementary Materials. Algorithm 1 gives an overview of the information flows exchanged by the model components, the computations executed by these components, and the interactions between the model and the environment.

The algorithm involves a first cycle (line 1) where each step corresponds to a card drawn from the deck. In each step of the cycle, the model first visually scans the deck card (line 2). Next, it processes the card features that correspond to the matching rule stored in the working memory and memorises these features for later use (lines 3–6; a non-neural memory is used for this purpose). A second nested loop allows the model to visually scan one target card after the other in each step (line 8). For each target card, the model reconstructs its features corresponding to the current selected matching rule (lines 9–12) and then compares them with those of the deck card stored in memory (line 13). When a target card matches the deck card, the model stops scanning the target cards and moves the deck card below the last scanned target card (line 14). The model then collects the resulting feedback (line 15). Finally, the model uses the feedback to update the working memory (lines 16).

Environment. The agent acts in a simulated environment comprising a square space of 100 × 100 pixels. The environment contains 'objects' (the cards) that the model can visually explore and move in space (see Fig. 6). The objects are cards representing polygons characterised by a unique combination of three visual properties (categories), where each has one of four possible attributes: colour (red, green, blue, or yellow), form (square, circle, triangle, or bar), and size (large, medium large, medium small, or small). This set of attributes generates $4^3 = 64$ combinations (cards). With respect to the original task, we substituted the 'number' category with the 'size' category because perceiving a different number of objects required higher resolution and this slowed the simulations. For the same reason, we also substituted the form attributes 'stars' and 'crosses' with the attributes 'squares' and 'bars', respectively.

Visual sensor. The visual sensor returns a 28 × 28 pixel RGB image covering a limited portion of the environment. The resulting 28 × 28 × 3 matrix is stored in a vector of 2352 elements that represents the input for the perceptual component. The visual sensor is first directed towards the deck card and then towards the target cards in sequence until the model finds a target card that matches the deck card.

Perceptual component. This component is implemented as a deep generative model, specifically a *deep belief network* (DBN; Hinton et al., 2006; Le Roux & Bengio, 2008) comprising two stacked *restricted Boltzmann machines* (RBMs; Hinton, 2012). In the following, we explain the main features of the component, but detailed descriptions of its functioning and learning processes are presented in the Supplementary Materials. An RBM is formed by two layers of units comprising a 'visible' layer and a 'hidden' layer, which are fully connected. A distinctive feature of RBM networks, and thus of the DBN, is that information can flow in both a bottom-up and top-down manner within it. The bottom-up flow of the network (from the visible layer to the hidden layer) reduces the dimensionality of the input pattern (Hinton & Salakhutdinov, 2006) and the top-down flow (from the hidden layer to the visible layer) produces a visible input. The capacity of the network to utilise the activation of the last hidden layer through a top-down information flow to produce the possible input that corresponds to this activation is an important property called *generativity* (Bengio, Goodfellow, & Courville, 2017). The perceptual component is trained offline with a novel algorithm (see Supplementary Materials), which allows it to extract the specific attributes of each card and represent them in a distributed manner, and thus the model can use the generativity to simulate top-down attention processes. For example (Fig. 6), we consider a case where the model perceives a 'large, red, triangle' deck-card and a 'medium large, red, square' target card, and the category selected at the higher levels is 'colour'. In this case, the model can lead to the activation of a red blob at the lower levels for each card in the sequence and decide that the two cards match. In Section S2.3 of the Supplementary Materials, we show that the bidirectional activation of the component can be repeated many times to simulate the activity reverberations in the visual working memory, thereby allowing us to study the possible loss of information in the presence of interfering distractors if stimulus–response delays are introduced. Given that the participants can freely observe the deck and target cards as many times as they like in the WCST, we assume that there is no loss of information while performing the visual matching of cards. As a consequence, we fixed the number of reverberations of the perceptual working memory to 1 cycle involving a single spreading bottom-up activation followed by a single top-down reconstruction.

Executive working memory. The perceptual component is formed by three units encoding the three matching rules (colour, form, and size). The activation of each unit encodes the likelihood that the corresponding behavioural rule is selected. In particular, the units have a continuous value ranging from 0 (low chance) to 1 (high chance), and they store, based on a recurrent self-connection, a representation of the possible matching rules to use. Activation is fuelled by the feedback signal with a binary value from {0, 1}. The feedback signal only affects the activation of the unit encoding the last selected rule, as follows:

$$m_{s,t} = (1 - \mu) \cdot m_{s,t-1} + \mu \cdot r, \quad (1)$$

Algorithm 1 Model: information flows, computations, and interaction loops with the environment.

```
for deckCard ∈ {1, 2, ..., 64} do
    (deckCardImage, deckCardPosition) ← VisualComponentScan(deckCard)
    attributePreactivation ← DBNForwardSpreading(deckCardImage)
    category ← SoftMax(workingMemoryState)
    attribute ← DisinhibitionOfCategoryAttributes(attributePreactivation, category)
    reconstructedDeckCard ← DBNGeneration(attribute)
    match ← False
    for targetCard ∈ {1, 2, 3, 4} AND match = False do
        (targetCardImage, targedCardPosition) ← VisualComponentScan(targetCard)
        attributePreactivation ← DBNForwardSpreading(targetCardImage)
        attribute ← DihinibitionOfCategoryAttributes(attributePreactivation, category)
        reconstructedTargCard ← DBNGeneration(attribute)
        match ← VisualComparision(reconstructedTargCard, reconstructedDeckCard)
    MotorComponentMoveDeckCard(targetCardPosition, deckCardPosition)
    feedback ← GetFeedback()
    workingMemoryState ← UpdateWorkingMemory(feedback)
```

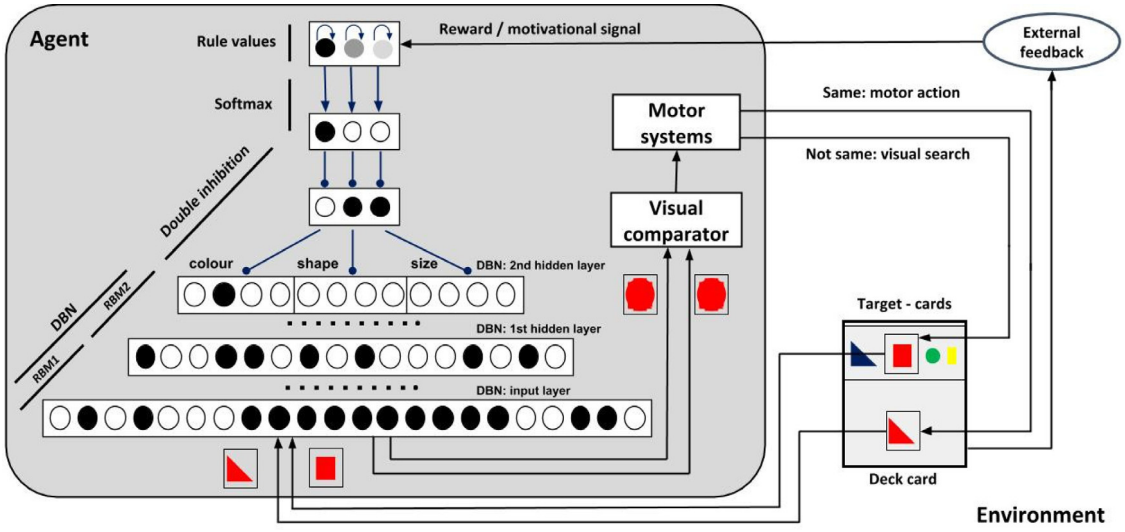

Fig. 6. Architecture of the model showing the deep belief network for perception, disinhibition mechanism for rule selection, and the rule values and softmax function for matching-rule selection. A stimulus used in the WCST is shown at the bottom right, where the small square frames around the red triangle and the red square represent two 100 × 100 pixel images corresponding respectively to a deck card and a target card collected by the system visual sensor in successive steps. The two analogous squared frames around the two red circles under the 'visual comparator' are the images obtained by considering the fact that the high levels of the model focus on the 'colour' category and the 'red' attribute to compare the two input cards. (For the reference to colours in this figure, the reader is referred to the web version of this article.).

where $m_{s,t}$ is the new activation for the rule unit, $s \in \{1, 2, 3\}$ is the index for the selected rule, $m_{s,t-1}$ is the previous activation of the unit, $(1 - \mu)$ is the strength of the unit recurrent connection, μ regulates the impact of the feedback on the memory, and r is the feedback signal, which is equal to 1 in the case of positive feedback (matching the deck and target cards) and 0 otherwise. In the case of positive feedback, the parameter μ assumes a fixed value of 0.7, whereas in the case of negative feedback, μ is considered to be a free parameter that possibly has different values (see Section 4.1.1 for details regarding the search for the model parameters). We used this approach because previous studies suggest that disengagement (switching after negative feedback) is a critical feature for detecting individual differences and also for pathological behaviours assessed with the WCST (Monchi et al., 2004; Zanolie et al., 2008). The parameter μ is the first of the three key parameters in the model investigated in the simulations.

All non-winning units of working memory decay exponentially towards a baseline value as follows:

$$m_{l,t} = (1 - \phi) \cdot m_{l,t-1} + \phi \cdot \alpha, \tag{2}$$

where $m_{l,t}$ is the value related to the losing unit l ($l \in \{1, 2, 3\}$; $l \neq s$) at time t, $1 - \phi$ is the strength of the recurrent connection, and α (set to 0.5) is the baseline value to which the memory unit activation converges. A high value of ϕ causes a high rate of information forgetting. The parameter ϕ is the second of the three key model parameters investigated in the present study.

Perceptual manipulator. This component implements the following three processes. The first process is a winner-take-all (WTA) competition that receives the values from the working memory as inputs and chooses the matching rule based on the *softmax* function:

$$Pr(k = s) = \frac{\exp(m_k / \tau)}{\sum_{q=1}^{3} \exp(m_q / \tau)}, \tag{3}$$

where $Pr(k = s)$ is the probability of the event that the matching rule k ($k \in \{1, 2, 3\}$) is selected ($k = s$) and τ is the 'temperature parameter' in the *softmax* function for regulating the randomness of the selection. A high value of τ leads to high randomness/exploration of the behavioural rules. The parameter τ is the third of the three key model parameters investigated

in the present study. The probabilities $Pr(\cdot)$ sum up to 1 and they are used to stochastically select the matching rule for use. It should be noted that the stochasticity of the softmax function is the unique source of the behavioural variability of the model. The second process leads the winning unit in the WTA competition to apply a double inhibition mechanism to disinhibit the units of the last DBN hidden layer corresponding to the chosen category. The third process is a localistic winner-take-all mechanism (Srivastava, Masci, Kazerounian, Gomez, & Schmidhuber, 2013) involving the units encoding the attributes of each category group, and it is applied to the last DBN hidden layer before disinhibition. In this process, the unit with the maximum sigmoid activation (e.g., encoding 'red') is assigned an activation value of 1, whereas the other units (e.g., encoding blue, green, and yellow) are assigned an activation value of 0. For example, this process can allow the activation of the 'red' attribute if the 'colour' category is disinhibited.

Visual comparator. This component computes the Euclidean distance between the two reconstructed images of the deck card and the focused target card. Using a fixed threshold β ($\beta = 0.1$), the component returns a Boolean value representing the result of the comparison ('match'/'not match'). This process is an abstraction of neural comparison processes (e.g., see Santucci, Baldassarre, & Mirolli, 2016).

Motor component. This component encompasses two mechanisms. The first mechanism receives the positions of the deck and target cards, and locates the visual sensor (saccades) on them in a sequential manner. This approach captures the essence of more sophisticated attentional mechanisms that the model could use in future studies, such as a bottom-up attention mechanism based on image salient areas or the inhibition-of-return mechanism (Klein, 2000) that we used in previous models (e.g., Baldassarre et al. 2019).

The second mechanism in the motor component receives the positions of the deck card and the matched target card, and performs a movement to bring the deck card close to the matched target card. This action is hardcoded in our method and it is assumed to be learned by the model before the test (see Baldassarre et al., 2019).

4. Results

The results are presented in the following three sections. In Section 4.1, we present a validation of the model by showing how it can reproduce the behaviour of healthy and pathological humans in the WCST. In Section 4.2, we explore the relationships between the model's behaviour and its key parameters. In particular, we present the results obtained by correlation analysis to investigate the links between the model's parameters and the behavioural indices exhibited by humans in the WCST. In this section, we also present the results of a 'lesion' experiment where the key model parameters were altered by setting them to extreme values to further investigate the links between these values and the behavioural results. Finally, in Section 4.3, we analyse the internal functioning of the three key elements of the mechanism used by the model for the internal manipulation of representations, thereby highlighting its key role in the production of flexible behaviour. The behaviour, underlying reasoning processes, and internal representations of the model can be observed in action in a video at: https://youtu.be/pnBWWqhULsE

Table 1
Values of the parameters in the models that obtained the best fits to the target WCST data related to the behavioural indices for healthy participants and frontal patients (data from Heaton et al., 2000), and for Parkinson controls and patients (data from Paolo et al., 1995).

	Error sensitivity (μ)	Forgetting speed (ϕ)	Distract-ibility (τ)
Healthy young participants	0.26	0.26	0.14
Frontal patients	0.05	0.47	0.14
Healthy old participants	0.16	0.47	0.14
Parkinson patients	0.05	0.37	0.17

4.1. Validation of the model with human data

We targeted four groups of participants to validate our model (Heaton et al., 2000; Paolo, Tröster, Axelrod, & Koller, 1995). All of the participants completed the standard version of the WCST (Heaton et al., 2000). In particular, we considered two pathological groups of 59 frontal patients with a local or diffused frontal lesion (average age of 42 $\pm$ 14.32 years), and one group of 181 Parkinson patients (average age of 68.92 $\pm$ 8.28 years). The pathological group and Parkinson group were paired by age, education, and IQ with control groups of 362 young adults and 162 old adults, respectively.

4.1.1. Model configurations that obtained the best fits to data from healthy and pathological humans

We searched for the values of the three key parameters in the model (μ: sensitivity to negative feedback errors; ϕ: working-memory forgetting speed; and τ: exploration/distractibility) using a brute force search algorithm (Van Geit, De Schutter, & Achard, 2008), as explained in detail in Section S4.1 of the Supplementary Materials. The large number of models tested with this technique also allowed us to use a *sensitivity analysis* (Hamby, 1994) to assess the performance of the obtained parameters. Table 1 shows the values of the model key parameters obtained with the automatic search method, that is, the model parameters that resulted in the lowest minimum squared error between the behavioural indices for the human groups and the model groups. Fig. 7 presents a general view of the model's parameter configurations that obtained the best fits to the four human populations. The plot shows the differences between young healthy participants and the other three populations, thereby supporting the idea that the effect of ageing on healthy old participants can mimic a frontal impairment (Dennis & Cabeza, 2012; Sullivan et al., 2001).

Among the models used to fit the data reported by Heaton et al. (2000), the 'pathological model' with the parameters that obtained the best fit to the data related to frontal patients had a lower μ, higher ϕ, and similar τ compared with the 'healthy model' that obtained the best fit to the control group (healthy young participants).

Among the models used to fit the data reported by Paolo et al. (1995), the model configurations fitted to the Parkinson patients had lower sensitivity to negative feedback (μ) and higher distractibility (τ) compared with the paired control group. Surprisingly, the Parkinson model had a low working memory forgetting value (ϕ) compared with the related control group (healthy old participants), although it was still higher than that for the healthy younger participants in the study by Heaton et al. (2000). Moreover, healthy old participants had a similar forgetting speed to frontal patients.

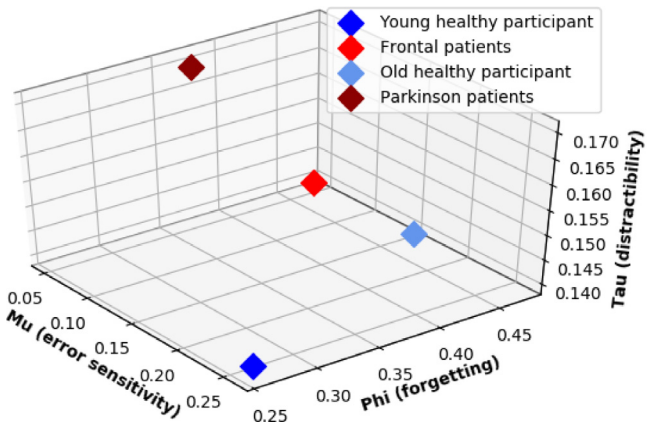

Fig. 7. Three-dimensional representations of the parameter configurations in the models that obtained the best fits to the four human populations. (For interpretation of the reference to colour in the figure legend, the reader is referred to the web version of the article.).

4.1.2. WCST indices for the healthy and pathological models, and corresponding human groups

We further validated the model versions with the parameter configurations discussed in the previous section by studying the accuracy with which they reproduced the multiple behavioural indices exhibited by the corresponding groups of human participants. For each model parameter set (human group), we ran and compared 59 simulated participants (that varied by using different seeds in the random number generator) with the 59 frontal participants, and 59 other simulated participants with the 362 healthy participants considered by Heaton et al. (2000). Fig. 8 shows the average values of the indices for the healthy humans and those obtained by the model, and Fig. 9 shows an analogous comparison for frontal patients. The comparisons showed that the model reproduced the values of the WCST indices for all of the human groups considered with high accuracy. Table 2 presents the p-values obtained from statistical comparisons of the behavioural indices produced by the models and those for the human groups. The indices were not statistically different ($p > 0.05$), except CC was higher in the two models compared with the human participants (healthy human versus healthy model: 5.18 ± 1.52 versus 5.9 ± 0.4, $p < .01$; pathological human versus pathological model: 3.46 ± 2.25 versus 4.6 ± 1.0, $p < .01$), and FMS was higher in the healthy model compared with the humans (1.4 ± 1.3 versus 0.67 ± 1.09, $p < .01$). These differences were due to the very low variability of the behaviour of the model (see the standard deviations in the figures) leading to a statistically disproportionate weighting on small mean differences. The lower variability of the data obtained by the model could have been caused by the simplicity of the architecture of the model compared with the human brain. In particular, the softmax function (Eq. (3)) is the unique source of variability in the model, whereas the human brain exhibits high variability in terms of its architecture and functioning (participants pursue multiple goals in parallel even when performing the task, e.g., they might aim to save energy or be socially compliant), thereby leading to individual differences in cognition and behaviour (Barch et al., 2013; Chen et al., 2015; Finn et al., 2015; Hearne, Mattingley, & Cocchi, 2016; Kanai & Rees, 2011).

In the experiments based on Parkinson patients and the paired healthy control group reported by Paolo et al. (1995), most of the WCST indices obtained by the model were not statistically different from those for the human groups. Again, statistical differences were found only for CC and FMS, where the values were higher using both models because of the same reasons explained above for frontal patients. The statistically non-significant t-tests did not indicate that the results were the same but they further corroborated the capacity of the model to reproduce multiple behaviours of the target human groups.

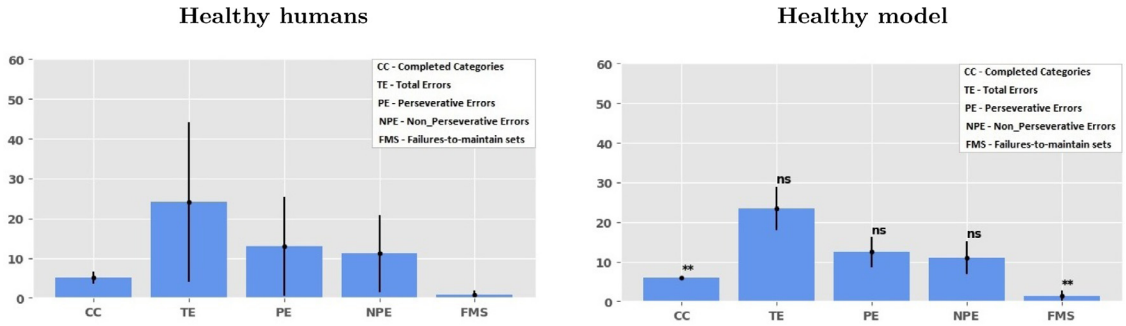

Fig. 8. Healthy condition: comparison between the healthy model group and healthy human group (** indicates a statistically significant difference at $p < 0.01$).

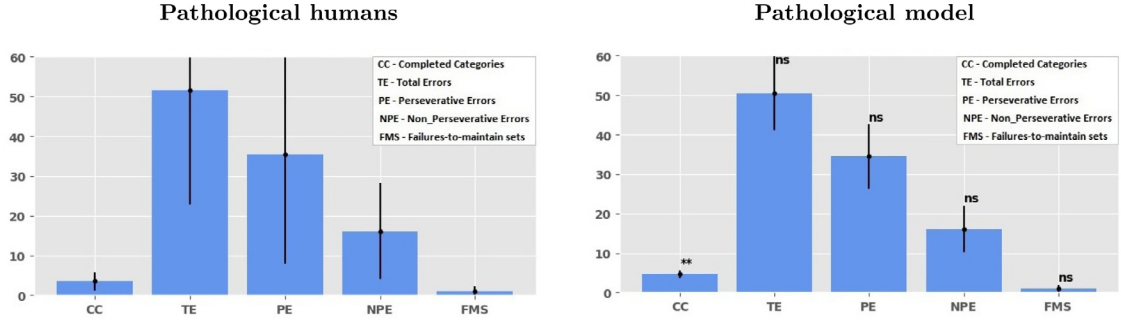

Fig. 9. Pathological condition: comparison between the artificial impaired group and human frontal patients (** indicates a statistically significant difference at $p < 0.01$).

Table 2
Statistical comparisons (p-values, two-tailed t-tests) of human data versus model data involving the healthy and pathological conditions (data from Heaton et al., 2000), and healthy and Parkinson conditions (data from Paolo et al., 1995). The statistically significant p values ($p < 0.05$) are highlighted in *Italics*.

Participants	Indices				
	CC	TE	PE	NPE	FMS
Healthy (Heaton et al., 2000)	*.001*	.800	.748	.920	*.001*
Frontal (Heaton et al., 2000)	*.001*	.784	.794	.953	.565
Healthy (Paolo et al., 1995)	*.004*	.873	.763	.969	*.003*
Parkinson (Paolo et al., 1995)	*.004*	.678	.635	.964	*.000*

Table 3
Pearson's r values indicating the correlations between the key parameters in the model (μ, ϕ, and τ) and the different WCST indices. Except for the correlation related to ϕ-FMS, all of the correlations were statistically significant ($p < 0.001$). Correlations stronger than |0.3| are highlighted in *Italics*.

Indices	Parameters		
	μ	ϕ	τ
CC	0.25	−0.14	*−0.67*
TE	*−0.60*	0.17	*0.34*
PE	*−0.58*	0.07	−0.05
NPE	−0.26	0.24	*0.75*
FMS	0.04	0.00	*0.73*

4.2. Study of the internal functioning of the model

4.2.1. Relationships between the key parameters of the model and WCST behavioural indices

We analysed the relationships between the three key model parameters and the WCST indices by considering their correlations measured using Pearson's r (Table 3). Figure S5 in the Supplementary Materials presents visualisations of the possible non-linear relationships between these values. The analysis showed that CC, indicating the global performance of the model, tended to correlate with high error sensitivity (μ; $r = +0.25$), low forgetting (ϕ; $r = -0.14$), and low distractibility (τ; $r = -0.67$). The analysis also indicated that TE had a negative relation with error sensitivity (μ; $r = -0.60$) and a positive relation with distractibility (τ; $r = +0.34$). The analysis also showed that PE had a robust negative relation with negative feedback processing (μ; $r = -0.58$) but negligible correlations with the other two parameters. The analysis also indicated that NPE had a remarkably positive correlation with distractibility and erratic behaviours (τ; $r = +0.75$), a moderate positive correlation with forgetting (ϕ; $r = +0.24$), and a negative correlation with the error sensitivity (μ; $r = -0.26$). Finally, the FMS errors had a strong correlation with distractibility (τ; $r = +0.73$) and negligible correlations with the other two parameters.

4.2.2. Effects of focused alterations of the model on WCST behavioural indices

The simulated lesion technique allowed us to further investigate the role of each single key model parameter in the production of the flexible behaviour measured using the WCST indices. In particular, Table 4 shows the three sets of parameters used to obtain the three alternative versions of the control model investigated in this study. The first model called the 'extreme perseverative model' (EPM) is characterised by a very low value for μ. The second model called the 'distracted model' (DM) is characterised by a very high value for τ. The third model called the 'irrational model' (IM) is characterised by a high value for ϕ.

A global view of the proportions of error with the three altered models (Fig. 10) confirmed that the EPM and DM had opposite PE/NPE imbalances, where the former had high PEs and low NPEs, and the latter had low PEs and high NPEs. Moreover, the EPM had few FMS errors whereas the DM had many. The IM had slightly more errors than the healthy model, with an imbalance towards NPEs.

Table 4
Parameter values used in the impaired models for producing focused alterations. Values in *italics* represent the altered parameters with respect to the values found by fitting the data of the healthy participants in the study by Heaton et al. (2000).

	μ	ϕ	τ
Control model	0.26	0.26	0.14
Extreme perseverative model	*0.001*	0.26	0.14
Distracted model	0.26	0.26	*0.4*
Irrational model	0.26	*1*	0.14

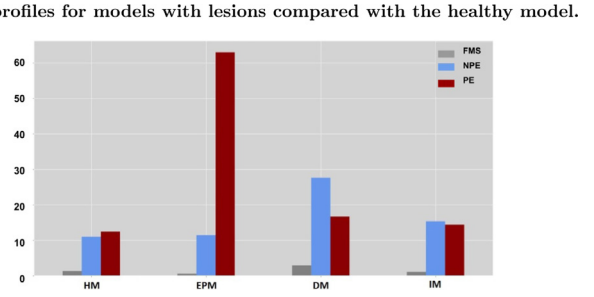

Fig. 10. Proportion of errors in the altered models compared with the healthy model. HM: healthy model; EPM: extreme perseverative model; DM: distracted model; IM; irrational model; PE: perseverative errors; NPE: non-perseverative errors; FMS: failure-to-maintain set errors. (For the colours of the card items, the reader is referred to the web version of the article.).

4.3. Analysis of the functioning of the model mechanism for the internal manipulation of representations

In this section, we describe our investigation of the relationships between the behaviour of the model and the computations of the core components of the three-component hypothesis instantiated in the model, that is, the executive working memory, top-down representation manipulator, and visual working memory. First, we studied the relationships between the activation of the executive working memory units and the resulting model actions and errors in healthy and pathological conditions. Next, we studied the internal functioning of the perceptual component, particularly to show how the top-down manipulator based on the sub-goals of the model affected its internal representations of the input stimuli.

4.3.1. Executive working memory dynamics

To illustrate the functioning of the working memory component, we plotted the working memory activations and related behavioural responses for the five models. In particular, we considered the healthy model and pathological model with the parameters obtained by fitting the data reported by Heaton et al. (2000) (Fig. 11), and the three altered versions of the model considered in the previous section, that is, EPM, DM, and IM (Fig. 12).

In the graphs, the parts of the curves increasing from 0.5 to 1.0 during ten correct responses indicate a successfully completed card category. Curves decreasing from 1 to 0.5 indicate that the desirability of a matching rule decreases, thereby possibly leading to the selection of a different rule. Low values in the two curves followed by the increase in the third curve indicate inferential reasoning by exclusion. Sequences of several small peaks above the baseline (0.5) after a category change suggest the failure of an inferential reasoning process. A stable horizontal curve coupled

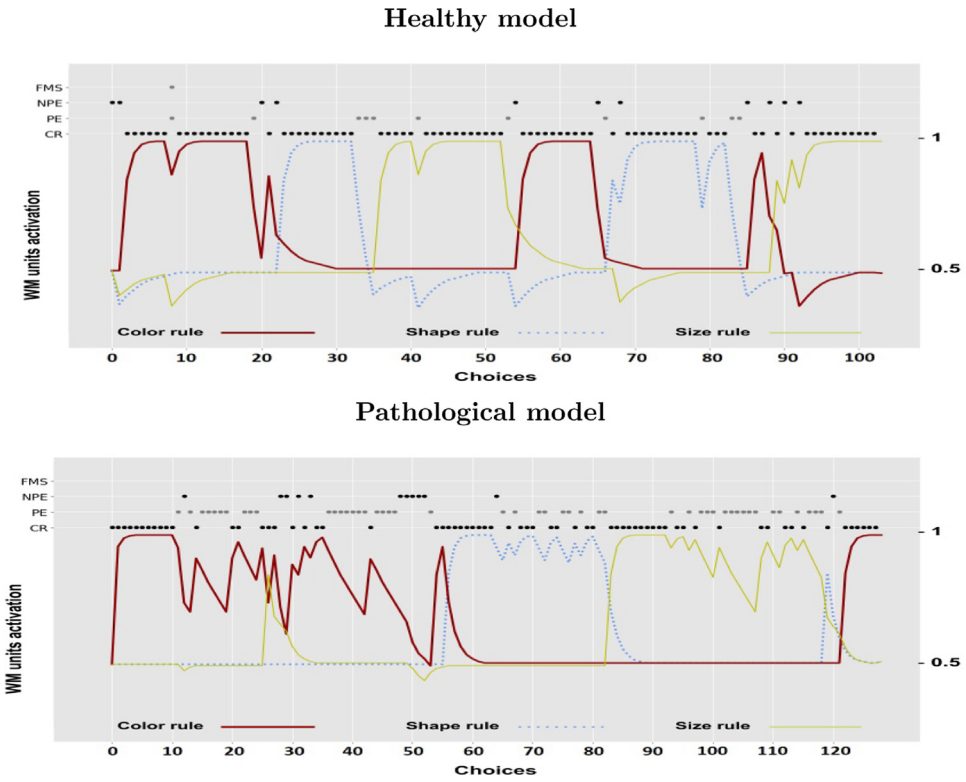

Fig. 11. Internal functioning of the executive working memory in the healthy model and pathological model. Each line represents the activation of a memory unit encoding a specific matching rule: thick red line: colour-based matching rule; dotted thin blue line: shape-based matching rule; and continuous yellow line: size-based matching rule. The dots at the tops of the graphs indicate single instances of correct responses (CR) or errors (PE, NPE, or FMS errors). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

with many errors corresponds to a strong perseverance tendency (e.g., see the graph for the EPM). Conversely, a graph with several increases and decreases in different curves indicates an erratic behaviour (e.g., see the graph for DM).

The healthy model (Fig. 11) completed the six WCST categories and performed correct reasoning after category changes, possibly after one or two errors and 'inference by exclusion'. Moreover, in the choice interval of 85–95, the model had many NPEs after choosing the colour rule and receiving positive feedback in trials 86 and 87. In this case, the correct sorting rule was size but the model chose a target card that shared both the colour and size attributes with the deck card. The model focused on the colour rule (the red solid line representing the colour priority increased after positive feedback) and the positive feedback led the model to increase the priority of the colour rule. Next, in trial 88, the deck card and target card shared the colour attribute but not the size attribute (the correct sorting rule was still size), so the model received negative feedback and it lowered the colour priority and chose the correct size rule.

The pathological model (Fig. 11) produced many prolonged incorrect activations that led to both PEs (e.g., in the choice intervals of 15–25 and 35–45) and NPEs (in the choice intervals of 27–33 and 47–53). Despite the presence of both error types, the model had two long series of PEs (choice intervals of 65–82 and 95–120) and it continued to choose the size rule after changing from size to colour.

The EPM (Fig. 12) obtained a similar trend to the pathological model but with more prolonged incorrect fixed choices (e.g., see the choice interval of 65–115). Interestingly, the inability to switch the sorting rule after negative feedback caused many small perseverative trends during the inferential reasoning process (e.g., see the choice interval of 30–40).

The DM (Fig. 12) had a high number of sudden random changes in working-memory activations, which caused many NPEs. Interestingly, the model also produced scattered PEs (see Section 5 for an explanation of this phenomenon). Moreover, due to the erratic behaviour, the model often chose an incorrect rule despite its low priority value, thereby lowering it further (e.g., see the choice interval of 54–59).

The IM (Fig. 12) produced an almost healthy-like plot, with the fundamental exception that many errors were produced when the model should have changed the matching rule. In this case, the priority values of all the rules immediately dropped to the same baseline, and the model did not keep track of the effects of past actions or prevent the execution of bad choices based on previous feedback. As a consequence, on average, the model had to make more choices to find the correct rule in a random manner, and thus it incurred some PEs and several NPEs.

4.3.2. Perceptual component: internal representations

After training (see Section S2.2 in Supplementary Materials), the DBN that implemented the perceptual component could extract the specific attributes of each card with its highest neuron layer while also generating the images corresponding to these attributes in the input layer by interacting with the top-down manipulator. The model used the latter capacity based on its *generativity* to compare the WCST cards in the selected attribute category (colour, form, and size). To investigate the quality of these representations, we analysed the images reconstructed by the component when single units from the first and second hidden layers in the network were manually activated in isolation (this operation simulated the disinhibition effect of the top-down manipulator when performing the WCST). Fig. 13 shows the images generated by activating the units in the first hidden layer (graphs on the left) or the units in the second hidden layer (graphs on the right). The figure shows that the images generated by activating single units in the first hidden layer involved

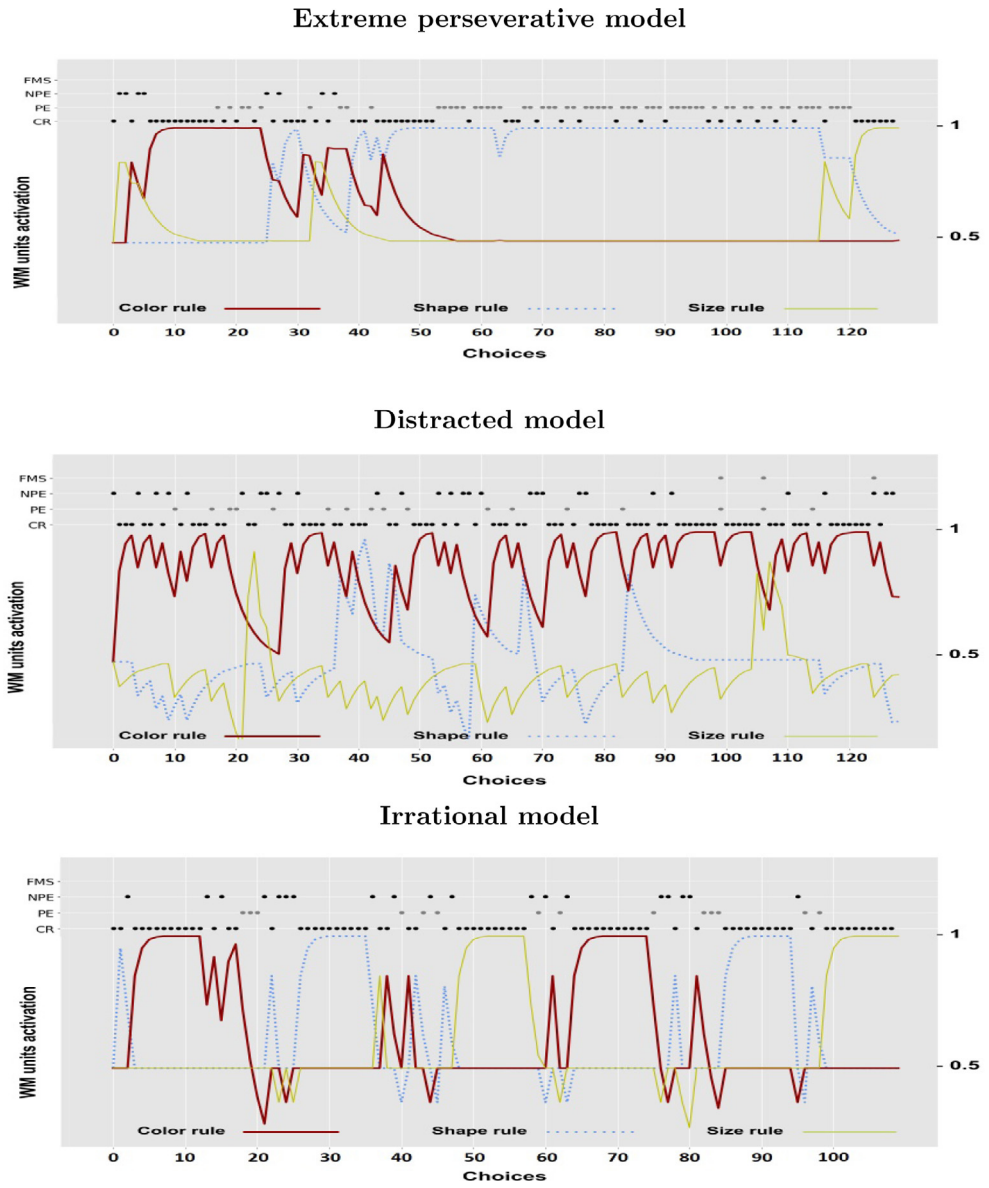

Fig. 12. Internal functioning of the executive working memory in the models with focused alterations. Each line represents the activation of a memory unit encoding a specific matching rule: thick red line: colour-based matching rule; dotted thin blue line: shape-based matching rule; and continuous yellow line: size-based matching rule. The dots at the top of the graphs indicate instances of correct responses (CR) or errors (PE, NPE, or FMS errors). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

different attributes of categories (e.g., mixed colours, forms, or sizes). By contrast, the images obtained by activating single units in the second hidden layer involved disentangled representations of each specific category attribute independently of the other attributes. For example, a unit encoded the 'prototype' of the blue colour independently of the size and shape of the object, and another unit encoded the prototype of the triangle shape independently of the colour and size of the object.

5. Discussion

5.1. Interpretations of the results

5.1.1. Cognitive profiles of the simulated participants

The results obtained by the parameter fitting procedure described in Section 4.1.1 (see Fig. 7 for an overview) showed that the cognitive profiles of healthy young participants were very different compared with those of the three groups of healthy old participants, frontal patients, and Parkinson patients, thereby supporting the idea that the effect of ageing on healthy old participants impairs the executive functions (Dennis & Cabeza, 2012; Sullivan et al., 2001). The 'pathological model' fitted to the frontal patients obtained different values for the μ and ϕ parameters and a similar τ value compared with the 'healthy model' fitted to the healthy young participants. These results suggest that frontal patients: (a) are less flexible at adapting their behaviour after negative feedback (μ); and (b) they have a lower capacity for remembering and reasoning about the appropriate behaviour to undertake based on experience (ϕ). These findings highlight the fact that frontal patients exhibit a mixture of deficits, with a tendency to perseverate in non-adaptive behaviours and poor executive functioning (e.g., see Barceló, 1999).

The model configurations fitted to Parkinson patients and healthy old participants obtained very different profiles. In particular, Parkinson patients exhibited less sensitivity to negative feedback (μ) compared with the paired control group. This difference might have been related to their altered capacity for

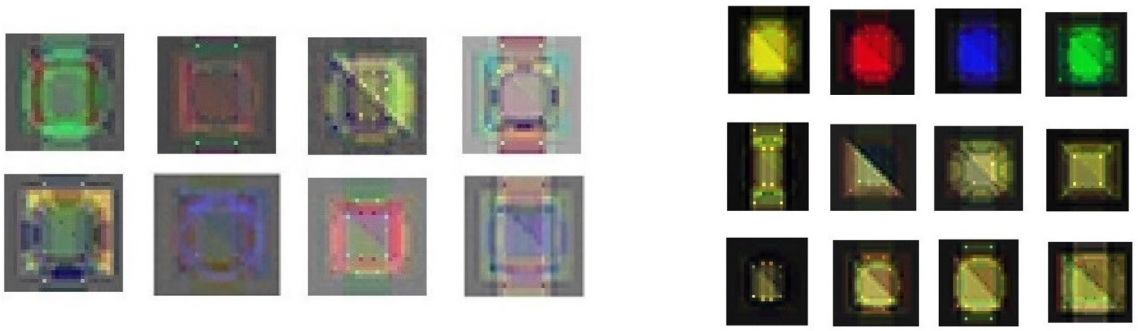

Fig. 13. *Left*: Images generated by activating a sample of single neurons in the first hidden layer to show how each encodes a mixture of colour, shape, and size attributes. *Right*: Images generated by activating single neurons in the second hidden layer to show how each image encodes a specific disentangled category attribute, which can be seen by considering that the three rows of graphs refer to the three categories (from top to bottom: colour, form, and size) and the four columns refer to different category attributes (colour: yellow, red, blue, and green; form: bar, triangle, circle, and square; size: small, medium small, medium large, and large). For the reference to colours in this figure, the reader is referred to the web version of this article.).

processing reward and feedback, which is a distinctive feature of the disease caused by the corruption of the dopamine system (Volpato et al., 2016). The comparison also indicated higher distractibility (τ), which might have been related to the lower capacity of Parkinson patients to 'lock-in' on the correct behaviour, and this is another relevant function of the dopamine system (Finn et al., 2014). The comparison also highlighted an unexpected result where the working-memory forgetting (ϕ) of Parkinson patients was low compared with the related control group (healthy old participants), although it was still higher than that of the healthy young participants in the study by Heaton et al. (2000). This result could be explained by the effects of Parkinson treatments on the executive role of the working memory, thereby possibly affecting the reasoning-by-exclusion process involved in the performance of the WCST (Fallon, Mattiesing, Muhammed, Manohar, & Husain, 2017).

A second unexpected result was that the working-memory forgetting speed (ϕ) of healthy old participants was similar to that of the frontal patient group members in the study by Heaton et al. (2000). This result has an interesting explanation, which was captured by the model, as follows. The frontal patient group had an average age of 42 ± 14.32 years whereas the healthy old participants group had an average age of 69.74 ± 6.96 years, and thus the latter group was probably affected by age-related weakening of the working memory (Daselaar, Cabeza, Ochsne, & Kosslyn, 2013).

5.1.2. Cognitive processes and behavioural responses

The results reported in Section 4.1.2 highlight the interesting relationships between cognitive processes and the behavioural indices scored in the WCST. For example, the positive correlation between CC (indicating global performance) and the error sensitivity (μ) supported the construct validity of the WCST, that is, the test evaluates the capacity to change the categorisation rule after negative feedback. Furthermore, in order to exhibit adequate performance in the WCST, a participant requires an intact working memory storage capacity (negative correlation between CC and the working memory forgetting speed parameter, ϕ) and attention abilities (negative correlation between CC and the distractibility parameter, τ).

The correlations between TE and the behavioural indices supported our considerations regarding the CC index (both the CC and TE indices indicate the global performance in the WCST). In particular, the negative correlation between TE and the error sensitivity parameter (μ), as well as its positive correlation with the distractibility parameter (τ), supported the idea that negative feedback reactivity and attention abilities (operationalised as 'distractibility' in this study) are key processes when solving the WCST.

The negative correlation between PEs and the error sensitivity parameter (μ) supported the findings of classic studies of the WCST, which associated perseverative rigid behaviours with difficulty in adapting behaviour after negative feedback (Dehaene & Changeux, 1991).

Interestingly, the positive correlations between NPE with distractibility (τ) and the working memory forgetting speed (ϕ) confirmed the previously claimed relationships between this type of error, the lack of attention abilities, and 'reason by exclusion' failures (Barceló & Knight, 2002; Dehaene & Changeux, 1991). Moreover, the negative correlation between NPEs and the error sensitivity parameter (μ) suggested that reasoning by exclusion (the failure of which tends to cause NPEs) strongly depends on the capacity to evaluate external feedback.

Finally, the strong correlation between the FMS errors and distractibility (τ) confirmed that maintaining correct behaviour is highly dependent on the ability to maintain internal focus on the selected categorisation rule. Overall, these results are in agreement with previous findings (Section 2) and the results obtained by different models (Section 5.3).

5.1.3. Brain lesions, cognitive deficits, and behavioural impairments

The results presented in Section 4.2.2 and summarised in Fig. 14 show the possible correspondences between model lesions and brain lesions, and the consequent behavioural impairments.

The EPM characterised by a very low error sensitivity (low μ parameter) had a high number of PEs. This lesion might correspond to malfunctioning of the ventral ACC involved in the motivational processing of errors (Lie, Specht, Marshall, & Fink, 2006). This structure together with medial and ventral cortical and sub-cortical areas regulates negative emotions (Etkin, Egner, & Kalisch, 2011) and the processing of the affective valence of stimuli (Roy, Shohamy, & Wager, 2012).

The DM characterised by high distractibility (high τ parameter) obtained the opposite behavioural profile compared with the EPM, that is, an index imbalance toward NPEs compared with PEs, although with only a minor difference. Moreover, the DM had an increased number of FMS errors, thereby confirming an unstable attention focus. This alteration might correspond to an impairment of the dorsal ACC that interacts with the dorsal and frontal cortices to influence decision making and response selection (Bush et al., 2002; Heilbronner & Hayden, 2016).

Finally, the IM characterised by a high working memory decay speed (high ϕ parameter) had a slight imbalance towards NPEs. This alteration caused the working memory component to have

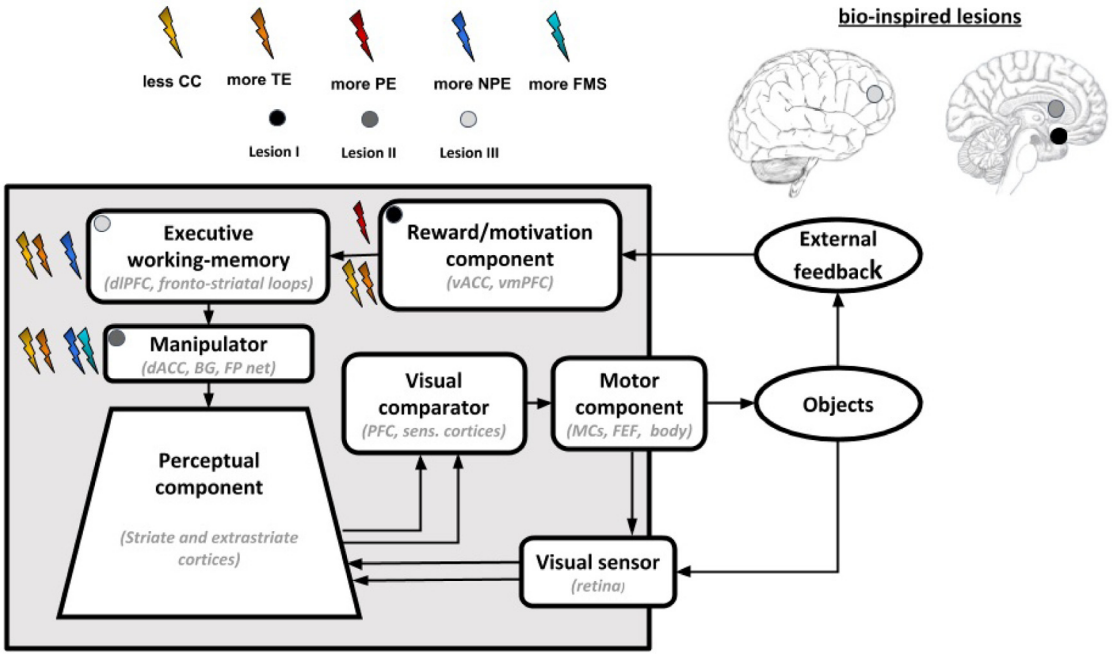

Fig. 14. Schema showing the architecture of the model and three 'focused lesions' obtained by altering specific parameters, which we applied to obtain three prototypical pathological conditions. Dots with a different intensity of grey represent the three alterations ('lesions') and the coloured bolts denote the decrease in performance/increase in errors that they caused during the performance of the WCST. As an example, lesion I mimicked effects analogous to those produced by a brain lesion in the ventral ACC and ventromedial PFC to impair the motivation system, and it caused a decrease in CC and increases in TEs and PEs. (For interpretation of the reference to colour in the figure legend, the reader is referred to the web version of the article.).

a high forgetting rate for previously chosen rules and it might correspond to an impairment of the brain system that supports executive working memory, particularly the dorsolateral PFC and its loops with basal ganglia (Mannella et al., 2013).

5.1.4. Working memory dynamics and consequent behaviour in the WCST

The results presented in Section 4.3.1 showed that different behavioural responses to the WCST corresponded to different executive working memory dynamics. The rule-based activation of the executive working memory of the model appeared qualitatively similar to those found in the neurons of the dorsolateral PFC of non-human primates performing variants of the WCST (Buschman, Denovellis, Diogo, Bullock, & Miller, 2012; Mansouri et al., 2006). These results indicate that executive working memory storage and updating are key processes when executing an adequate internal manipulation of representations, and thus they support flexible behaviour. For example, the healthy model fitted to the young healthy participants obtained overall good performance. However, it still exhibited exploratory behaviours supported by unstable activation of working-memory sub-goals. These behaviours might appear pathological. This phenomenon highlights the fact that the global behaviour of a participant can exhibit occasional cognitive failures.

The pathological model fitted to the data for frontal patients exhibited both perseverative behaviour (many PEs) and reasoning failures (many NPEs), thereby supporting the idea that frontal patients can be affected by different cognitive deficits beyond behavioural rigidity, such as working memory impairments characterised by inferential reasoning failures. Moreover, PEs are also exacerbated by a specific feature of Heaton's version of the WCST where the deck and target cards sometimes have more than one attribute in common, which can produce positive feedback regardless of whether the participant performs sorting based on the wrong matching rule (Dehaene & Changeux, 1991).

The EPM exhibited similar behaviour to the pathological model but it was characterised by a more severe insensitivity to feedback, thereby resulting in a higher number of PEs compared with the previous pathological model. The EPM was also more strongly affected by the feature of Heaton's version of the WCST described above than the pathological model. These dynamics support the relationship between cognitive rigidity involving feedback-independent maintenance of the same specific sorting rule and perseverative behaviour.

The DM exhibited erratic behaviour caused by severe impairment of the decision-making processes. In particular, this model produced a 'stimulus-driven behaviour', which was dissociated from the rule priority values, and it yielded a response based on one of the random specific attributes suggested by the input (colour, shape, or size). This behaviour resulted in the model frequently choosing a strategy with low desirability at a high level, thereby obtaining the lowest values for its working memory units compared with the other versions of the model. These impaired dynamics highlight the importance of attentional focus during the WCST because its deficit can cause unstable behaviour.

Finally, the IM exhibited healthy behaviour but with a highly impaired capacity for reasoning by exclusion, as shown by the fact that when the rule changed, the model required many attempts to discover the new rule. As found in experiments with human participants, it should be noted that the simulated experimenter represented by a software routine detected the errors but had no access to the decision-making processes of the model. This feature is a potential limitation of this version of the WCST due to many different cognitive factors such as an erratic decision-making process rather than repeated intentional wrong rule selection resulting in PEs. This limitation makes it more difficult to interpret the link between this behavioural index and the underlying cognitive processes.

5.2. Main theoretical contributions

The aim of this study was to explore a novel theoretical hypothesis regarding the cognitive processes that support flexible behaviour. In particular, the proposed hypothesis states that flexible cognition depends on the top-down manipulation of internal low-level perceptual representations. These representations can then support a suitable sensory–motor interaction with the environment in order to perform the task. This hypothesis agrees with empirical evidence obtained from studies of the brain regarding how attention and imagination processes involve low-level cortical areas (Baldauf & Desimone, 2014; Fuster & Bressler, 2015; Gazzaley & Nobre, 2012; Kosslyn, 1999; Mechelli et al., 2004). Importantly, the theory is supported by the fact that the PFC is strongly connected with the parietal cortex and this then forms a highly integrated system with motor areas (Passingham & Wise, 2012; Rizzolatti & Craighero, 2004). Functionally, this system plays an important role in representing 'affordances', which encode the conditions where actions can be successfully executed (Baldassarre et al., 2019; Fagg & Arbib, 1998) and it plays a key role in triggering actions (Jeannerod et al., 1995; Thill et al., 2013).

In our second and main contribution, the hypothesis was operationalised with a computational model. The tests showed that the model could perform the WCST. The core mechanism in the model allows a behavioural rule (goal) selected within the executive working memory to apply a top-down bias on the lower perceptual levels. This bias leads to a representation of the input that reflects the selected rule. The model diverges from previous models of the WCST (see the following section for further details) that directly link the selection of behavioural rules to the selection of actions. Instead, the model selects the manner in which the inputs are internally represented and these goal-biased representations then trigger suitable actions. Thus, the model represents a new tool for quantitatively studying the proposed hypothesis. This possibility was demonstrated in the present study by validating the model with data from multiple WCST experiments involving healthy young and old participants, as well as frontal and Parkinson patients, which have often been reproduced in isolation using previous models (Table 2 and Fig. 9).

Qualitative analysis of the internal representations of the executive working memory of the model demonstrated the key role that the internal manipulation of representations can play in flexible behaviour. In particular, this manipulation allowed the model to focus on the correct rule (sub-goal), and thus to perceive the cards in a 'rule-biased manner' that was suitable for supporting the correct responses (Figs. 11 and 12). The activations of the executive working memory of the model are compatible with those found in the PFC during the performance of the WCST (Buschman et al., 2012; Mansouri et al., 2006). Furthermore, the presence of perceptual representations with different levels of abstraction within the first and second hidden layers of the perceptual component reflects hierarchical information processing in the perceptual cortices of the brain (Baldassarre, Caligiore et al., 2013; Felleman & Van Essen, 1991; Mechelli et al., 2004). For example, the neurons in the primary visual cortex extract several low-level visual features from retina images (Rentzeperis, Nikolaev, Kiper, & van Leeuwen, 2014) whereas the neurons in the higher-order visual cortices tend to respond to macroscopic aspects of objects (Bracci, Ritchie, & de Beeck, 2017; Folstein, Palmeri, Van Gulick, & Gauthier, 2015).

5.3. Comparisons with other computational models

In this section, we present comparisons of our model with previously proposed computational models for studying the cognitive processes and neural mechanisms that underlie the performance of the WCST. Moreover, we consider other models that have not been used for studying the WCST but that have been employed for investigating executive functions and proposing hypotheses regarding the cognitive processes and biological mechanisms that support flexible cognition.

5.3.1. Models of WCST

Table 5 summarises the main features of the computational models used to investigate the WSCT, including our model.

Levine and Prueitt (1989) proposed a model for performing the WCST based on adaptive resonance theory (Carpenter & Grossberg, 1987). This model suggests that categorisation in the brain is based on an interactive relationship between top-down processes (e.g., expectations) and bottom-up processes (sensory information). The model qualitatively reproduces both perseveration and the novelty dependence of frontal patients. These behaviours are linked to the impaired integration of frontal structures that support both cognitive processes (attention to specific rules) and motivational processes (past effects of decision-related rewards and punishments). By contrast, our model supports the idea that a corrupted link between feedback computations (past rewards and punishments) and attention selection (selection of a specific rule) is caused by an impaired rule selection process (see DM in Section 4). In our model, this corruption produced slightly more PEs and many more NPEs, which were not considered by the authors.

Dehaene and Changeux (1991) proposed a model of the WCST that encompasses the top-down selection of rules and their integration with percepts, and a reward signal to select actions. This model also considers an 'intention layer' linked to the choice of the four target cards. The model reproduces PEs and a worsening of 'single-trial learning' as an index for measuring the length of a successfully completed series, which was not considered in the present study. These two results are based on impaired feedback processing and rule-based memory corruption. This model was not proposed recently but it incorporates various possible explanations of synaptic and molecular processes as the basis of solving the WCST. These processes are simulated at an abstract level, and they are related to reward processing and synaptic plasticity. Overall, this previous study provided an extensive functional analysis of the task, but it mostly focused on the perseverative tendency of patients in the WCST. In contrast to our proposed model, this previous model fails to analyse other types of errors, such as NPEs and FMS errors, thereby preventing the possibility of effectively discerning patient sub-populations, as achieved in our study, particularly determining the heterogeneous deficit profiles related to distractibility and perseveration.

Berdia and Metz (1998) proposed a model that simulates neural noise and synaptic instability based on two parameters comprising 'noise' and 'gain', and they linked them to the poor performance of schizophrenic patients in the WCST. This is one of the first models of the WCST to highlight the idea (as supported by our model) that a decrease in the influence of motivation (rewards/punishments) on behaviour can increase NPEs, thereby explaining poor global performance (e.g., in schizophrenic patients). The model considers attention processes related to the competition between categories, and reproduces PEs and NPEs in normal and schizophrenic participants. However, this model does not consider FMS errors (a sub-set of NPEs), which we included in our model. This omission prevents the investigation of multiple causes of NPEs, and particularly FMS errors. In our study, we found that NPEs can be caused by a reasoning-by-exclusion failure (IM) or by an unstable attention focus (DM), but only the latter type of failure caused a high number of FMS errors.

Amos (2000) proposed a model of the WCST that reproduces cortico-striatal loops and the related involvement of dopamine. In particular, their model suggests that the frontal cortex stores

Table 5
Overview of the main features of computational models used to investigate the WCST. 'Biological constraints' indicates whether the model incorporates fine-grained neural details (i.e., bio-constrained neuron models and detailed micro circuit connectivity; the other models, as ours, capture only the interactions between the brain macro-systems underlying the WCST). 'Data fitted' indicates whether the model was used to fit human experimental data (e.g., behavioural indices obtained during the solution of WCST), and the number in brackets indicates how many different data sets were used.

Models of WCST	Functions/Computational elements					Biological constraints	Data fitted	Number of free parameters
	Working Memory	Rule selection	Feedback computation	Sensory-motor processes	Top-down manipulation			
Levine and Prueitt (1989)	✓	✓	✓	✗	✗	✗	✗	1
Dehaene and Changeux (1991)	✓	✓	✓	✓	✗	✗	✗	3
Berdia and Metz (1998)	✓	✓	✓	✗	✗	✓	✓(2)	2
Amos (2000)	✓	✓	✓	✗	✗	✓	✓(6)	4
Kaplan et al. (2006)	✓	✓	✓	✗	✗	✗	✓(2)	2
Bishara et al. (2010) and Steinke, Lange, Seer, and Kopp (2018)	✓	✓	✓	✗	✗	✗	✓(7)	4
Caso and Cooper (2017, 2020)	✓	✓	✓	✓	✗	✓	✗	4
Steinke, Lange, Kopp (2020) and Steinke, Lange, Seer et al. (2020b)	✓	✓	✓	✓	✗	✗	✓(4)	8
Granato and Baldassarre (2020)	✓	✓	✓	✓	✓	✗	✓(4)	3

and selects the behavioural rule to follow, and that the striatum selects the target card based on the input card. By altering the two parameters linked to these key components, the model could fit the global performance (CC and TE) and the PEs of three groups of patients and related control groups. In particular, the model proposes that schizophrenic patients are affected by frontal impairment whereas Parkinson patients are affected by striatum deterioration. The model reproduces the behaviour of many human groups but it does not include non-perseverative and FMS errors, which are important for discerning many types of brain lesions. Furthermore, their model suggests that Parkinson patients are defined by a specific sub-cortical impairment, but our model showed that Parkinson patients can be characterised better by a heterogeneous impairment profile. In particular, Parkinson patients exhibit deficits involving both error sensitivity related to frontal-ventral impairment and distractibility related to alterations of both dorsal regions (e.g., dorsal ACC) and ventral regions (e.g., striatum).

Kaplan et al. (2006) reproduced the WCST with a model that uses a *Hamming network* (a feed-forward neural network for solving pattern recognition problems; Lippmann, 1987) to generate new strategies and a *Hopfield model* (an associative neural network; Hopfield, 1982) for storing them. These networks were used to reproduce both perseverative and failure-to-maintain errors in healthy and prefrontal patients. In this model, it is assumed that the former are caused by rigidity and the latter by attention failures. Similar to other models, this model does not consider NPEs, which we linked to both attention failures and failures of inferential reasoning in the present study. Furthermore, this model does not consider that attentional failures can cause PEs, which was shown by our model.

Bishara et al. (2010) proposed a model for performing the WCST and investigating the cognitive profiles of patients with substance addiction, schizophrenia (Cella, Bishara, Medin, Swan, Reeder, & Wykes, 2014), bipolar disorder (Cella et al., 2014; Farreny et al., 2016), and Parkinson patients (Steinke et al., 2018). The model encompasses an abstract component for computing positive feedback, negative feedback, and a 'choice consistence' (attention focus), as well as two different parameters for regulating the sensitivity to negative and positive feedback based on neuroscientific research that demonstrated a dissociation between the two (Monchi et al., 2004). Despite this evidence, Steinke et al. (2018) applied the model to study Parkinson patients and in agreement with our results, found that a single parameter for modulating the response to negative feedback was sufficient to fit their performance. This previous model can fit data related to a higher number of human groups but it does not consider all of the WCST behavioural indices, as included in our model.

Caso and Cooper (2017, 2020) proposed a computational model of healthy and Parkinson participants performing the WCST. This model aims to operationalise the 'schema theory' proposed by Schmidt (1976) by using a model architecture based on the neuroanatomy of basal ganglia and corticothalamic loops. This model highlights the important function played by basal ganglia as a fundamental 'selection machine' in the brain Redgrave et al. (1999); this function is also incorporated in our model. Moreover, similar to the model described next, this model stresses the idea that both motor selection (specific target cards) and 'conceptual selection' (sorting rule) influence the performance of participants during the solution of the WCST. It was concluded by Caso and Cooper (2020) that Parkinson patients exhibit a perseverative profile with a strong memory of past feedback, which corresponds to the lower 'memory decay' in our study. In addition, our model indicates that Parkinson patients exhibit high distractibility with respect to the correct strategy to follow, thereby supporting the idea that Parkinson patients exhibit a mixed impaired cognitive profile.

Steinke, Lange, Kopp (2020) proposed a model of the WCST that depends on the concept of two-level reinforcement learning. In particular, the model suggests that the trial-by-trial behaviour of participants is supported by model-based learning involving a decision-making process based on the sorting rule to choose and model-free learning based on the motor response executed after feedback (i.e., one of four target cards). The model was used to fit data related to healthy young participants and to show the existence of a perseverative tendency caused by response avoidance after negative feedback. This model was further validated by Steinke, Lange, Seer et al. (2020b) who fitted two groups of Parkinson patients ('on' and 'off' medication) and a matched healthy control group. The model demonstrated

that Parkinson patients had lower sensory-motor competencies (model-free learning processes) and cognitive deficits (model-based learning processes), and that the medication caused further cognitive symptoms such as cognitive inflexibility and attentional failures. Our model is comparable to the 'model-based learning' model version (only rule-based action) and it does not encompass a model-free learning component (card-based action). This previous model supported a card-specific effect but it was shown that the model-based choices could have a greater weight and fit better to the behaviour of some participants. Moreover, it was shown that the model with both model-based and model-free learning modalities produced the best fitted results, but it also had high complexity (eight or seven parameters). Our model can account for NPEs and PEs, and it can fit a comparable number of human populations based on only three free parameters.

Table 5 summarises the features of the models considered. Some of these models have objectives that go beyond the investigation of the WCST. For example, Caso and Cooper (2017, 2020) aimed to test schema theory, which was then operationalised in a model tested with the WCST. Similar to this approach, our principal objective was to investigate the three-component hypothesis based on the mechanisms that underlie flexible cognition and to support the theory by operationalising it in a model validated with WCST data.

We have presented qualitative comparisons of our model and other models, but we might perform quantitative comparisons to obtain more informative outcomes in future research, in a similar manner to that conducted by Steinke, Lange, Kopp (2020). Our qualitative comparison mainly indicates that none of the previously proposed models is supported by the manipulation of perceptual representations within low-level perceptual areas, and that category-based input representations might play relevant roles in the performance of the WCST. Moreover, some models involve sensory-motor components but in contrast to our model, none of them includes a visual search process that works together with a top-down manipulation mechanism to support flexible behaviour. In particular, most of the models assume the existence of hardwired semi-localistic representations of input patterns (e.g., representations based on one-hot vectors). By contrast, our model generates input representations based on a visual abstraction process applied to the raw visual input patterns. According to this analysis, most models of the WCST are based on a direct sequence of processes, as highlighted in Fig. 4, which comprises 'rule decision - selection of sub-goal - action performance - feedback computation'. By contrast, our hypothesis and model assume that the high-level decision-making processes related to the sorting rule to follow have a 'backward effect' on the internal low-level representations of stimuli, and these representations then affect action selection.

5.3.2. Models of executive functions

Table 6 summarises the features of some models proposed for investigating executive functioning and category learning in healthy and pathological participants, which are relevant to the issues investigated in this study.

Ashby et al. (1998) proposed a relevant model called 'COVIS' that represents both a theoretical framework and a computational implementation of neural structures that support category learning processes. This model is based on the hybridisation of symbolic and sub-symbolic mechanisms. In particular, the model is based on the idea that the brain performs category learning through: (a) a procedural implicit system that supports automatic action execution by mostly involving subcortical structures such as basal ganglia; and (b) a logical explicit system that supports rule-based behaviour by mostly involving cortical areas. This model assumes that the manipulation of response locations interferes with the procedural implicit system but not with the rule-based decision-making process. It was concluded that the WCST is mostly supported by a rule-based information-processing system, which is an idea that is also implemented in our model and other recently proposed models. The model was validated with WCST data collected from healthy participants and Parkinson patients (Hélie, Paul, & Ashby, 2012), and it highlighted the role of dopamine shortages in the executive deficits exhibited by these patients.

Monchi et al. (2000) proposed a biologically plausible model of working memory activation during the solution of two tasks comprising the delayed response task and WCST. In particular, the architecture of the model is based on biologically plausible neurons and it emulates the brain system formed by basal ganglia thalamocortical loops and working memory. The model was also lesioned to investigate working memory deficits in Parkinson and schizophrenia patients, and it predicted that the working memory deficits in Parkinson patients are caused by impaired disinhibition affecting the encoding and storage capacities of specific features. However, the model predicted that the working memory deficits in schizophrenia patients are related to the capacity for selecting specific features to store. Interestingly, our results support the possibility of rule selection impairment in Parkinson patients. In addition, we found that in addition to perseverative behaviour, incorrect selections can be caused by a working memory impairment (overlaps between rule representations) or by an altered top-down selection process that must evaluate the priorities of rules, where only the latter process is impaired in Parkinson patients.

Gilbert and Shallice (2002) propose a simple model of inhibitory control when performing the Stroop task. In particular, the model has two channels comprising a verbal channel and a colour channel, which compete to guide the behavioural response. Furthermore, this model has a 'task demand component' that stores the task requests ('name of the letters in input' or 'name of the colour in input') and a 'top-down control' component that indicates which task to follow. The model assumes that the switching costs between the two task demands are linked to top-down control and not only to automatic processes for response/conflict resolution. Interestingly, this model shares some features with our model but in a simpler form. In particular, similar to our model, this model performs top-down manipulation of the input representations by executing top-down selection of the input visual features to focus on the letters or the colours, and by implementing an 'intra-category competition' between the observed attributes to activate a specific attribute (e.g., red, green, or blue colours). These similarities with our model demonstrate that cognitive flexibility requires computational components linked to executive functions, particularly inhibitory control and working memory. Moreover, this model supports the idea that executive functions are based on the internal manipulation of representations.

Rougier et al. (2005) proposed a model that was used to reproduce the results of the WCST and it shares features with the other biologically grounded models. The model was implemented within the 'Leabra' framework (O'Reilly & Munakata, 2000) and it comprised various neural maps corresponding to the input layers, parietal cortices, prefrontal cortices, and motor output layers. The model was tested with the WCST and it reproduced the production of PEs when the PFC layer was lesioned, but other types of errors were not considered. In addition to the WCST, this model supports the idea that flexible behaviour is associated with the emergence of distributed rule-like representations, as also shown in the present study.

Another related model (the 'PBWM model') proposed by O'Reilly and Frank (2006) and updated by Hazy et al. (2007)

Table 6
Overview of computational models proposed to investigate executive functions and brain networks relevant to the issues investigated in the present study.

Models of executive functions	Functions/Computational elements					Biological constraints
	Working memory	Rule selection	Feedback computation	Sensory-motor processes	Top-down manipulation	
Ashby et al. (1998)	✓	✓	✓	✗	✗	✓
Monchi et al. (2000)	✓	✓	✓	✗	✗	✓
Gilbert and Shallice (2002)	✓	✓	✓	✗	✓	✗
Rougier et al. (2005)	✓	✓	✓	✓	✗	✓
Hazy et al. (2007), Kriete et al. (2013) and O'Reilly and Frank (2006)	✓	✓	✓	✓	✗	✓
Rigotti et al. (2010)	✓	✓	✓	✗	✗	✓
Granato and Baldassarre (2020)	✓	✓	✓	✓	✓	✗

and Kriete et al. (2013) replicated various functions of working memory by using an actor–critic model architecture (Sutton, Barto, et al., 1998) to reproduce the functions and macro-anatomy of the basal ganglia. In particular, this model was used to demonstrate the fundamental role of 'gating units', which are possibly used by basal ganglia to perform the uploading, storage, and download of information in working memory. This function is abstracted in our model by the winner-take-all competition involving the working memory units, and the basal ganglia disinhibition mechanism is used by the manipulator to allow the working memory to select lower-level perceptual representations.

Finally, Rigotti et al. (2010) proposed a recurrent neural network that allows rule selectivity in a version of WCST involving only 'form' and 'colour' categories. The model reproduced the internal neuronal dynamics involving mixed selectivity neurons, thereby suggesting that randomly connected neurons spontaneously exhibit mixed selectivity. The model did not assume an architecture for performing the WCST but instead it focused on the internal processes needed to solve context-dependent tasks with the aim of investigating how integrated neural networks can support the selection and storage of behavioural strategies.

5.4. Limitations and future work

Despite the contributions highlighted in the previous sections, the current model has some limitations that could be addressed in future work. In terms of validating the model with empirical data, a future study might aim to investigate the complexity of the model with respect to its parameters. In particular, as shown by Steinke, Lange, Kopp (2020), we could produce many versions of the model with different numbers of free parameters and compare them based on various indices that consider both the fitting accuracy and complexity of the model, such as the *Bayesian information criterion* (Schwarz et al., 1978).

A specific aspect of the architecture of the model that needs to be improved is the information flow between its components. A general strategy to address this issues could involve the use of deep neural networks (Bengio et al., 2017), as applied to the visual component (for details of this strategy, see the studies by Naselaris et al., 2018). Another possibly complementary strategy could involve grounding the information flows in the model by using a wholly neural dynamical system that mimics the macro-structure of the relevant brain components and that does not require a coded algorithm to control the information flows between the components of the architecture (e.g., as applied by Baldassarre, Mannella et al., 2013b, and by Mannella & Baldassarre, 2015).

Among the components of the model, a first limitation involves the simplicity of the executive working memory component, which comprises a few neural units for encoding the possible matching rules. This component might be improved by using mechanisms employed in other models of working memory, thereby enhancing the biological plausibility of the model (e.g., O'Reilly & Frank, 2006; Rigotti et al., 2010), or those used in deep neural networks (e.g., Hochreiter & Schmidhuber, 1997). Furthermore, the model can support an inferential process (i.e., reasoning-by-exclusion) but it cannot execute 'one-shot second-order inference'. For example, in the cases where (a) the model focuses on the colour feature, (b) the deck card and the target card share the colour and shape attributes, and (c) the model receives negative feedback, it decreases the priority value of the rule on which it is focusing (colour) but not that for the shape feature that is also not correct (in this case, the unique possible correct rule is the attribute not shared between the two cards, i.e., the number). In the future, this limitation could be addressed by implementing an internal reasoning process that considers both the specific feature on which the model focuses and by a further internal simulation of the potential feedback that would be obtained by alternative responses.

Another component of the model that could be enhanced is the overt attentional system, which currently depends only on a bottom-up attention process that allows the model to explore all stimuli in a stereotyped sequential manner. This approach is sufficient to study the WCST but this component might be improved by adding a top-down attention process for the goal-directed exploration of the elements in the environment, thereby improving the sensory–motor processes in the model and allowing it to perform other tasks (e.g., see Ognibene & Baldassarre, 2015; Sperati & Baldassarre, 2018).

Another important aspect of the model that should be enhanced is the process employed to acquire category and attribute representations. The current model uses a supervised learning algorithm, where the supervision is conducted by an unspecified external mechanism, e.g., other agents. Social learning might be important for the acquisition of categories, but we consider that most category learning by humans is derived from direct experience in the environment. Thus, in the future, we aim to employ other algorithms, particularly reinforcement learning algorithms (Caligiore et al., 2019; Sutton et al., 1998), to support the autonomous learning of categories by the model based on feedback from the environment following the performance of actions. The embodied elements of the model that support direct interactions with the environment are crucial for this purpose, for example the model might acquire categories and attributes based on the solution of multiple tasks, and thus the model could be used to address other neuro-psychological tasks.

A further improvement involves the key mechanism in the model related to the internal manipulation of representations. The current mechanism operates only on the highest level of the DBN used to implement the visual hierarchy. An important enhancement of the model would be a manipulator that can operate at multiple levels of abstraction within the hierarchy rather than at only one level as now. This would allow the model to *directly* manipulate representations with different levels of detail, as required by the downstream processes. For example, the

manipulator could focus on object attributes (e.g., 'circular shape') but also on their lower level features (e.g., different edges and corners of shapes), or on the relationships between the objects forming a scene (the latter would require the addition of more abstract layers to the model). Similarly, it could focus on the sub-goals that form the overall goals (Baldassarre et al., 2019), on the single actions that form the action sequences needed to achieve goals (Chersi, Mirolli, Pezzulo, & Baldassarre, 2013), or on predictions related to the consequences of actions (Baldassarre, Mannella et al., 2013b).

Overall, we consider that if developed in a suitable manner as described above, the model could show how the mechanism responsible for the internal manipulation of representations can support multifaceted manifestations of human cognitive flexibility, for example in reasoning and planning.

6. Conclusions

In this study, we proposed a new hypothesis to explain flexible cognition, as manifested by human participants performing the WCST. The hypothesis posits that flexible cognition depends on goal-directed processes that manipulate low-level perceptual representations, which then support the production of suitable actions. This hypothesis is novel compared with previously proposed methods and models of the WCST, which are all based on decision-making mechanisms that directly support action selection.

The hypothesis was operationalised by realising a computational model. This model depends on three main processes, which we assume are supported by specific brain systems. The first process, which involves the executive working memory that depends on the brain PFC and ventromedial basal ganglia, stores goals and behavioural rules. The second process, which involves perceptual working memories that depend on hierarchies of perceptual cortical systems, can extract and retain information at different levels of abstraction and generate lower representations based on the activation of patterns encoded in the higher levels. The third process, which involves the internal manipulation of perceptual representations by the brain system comprising the frontoparietal cortices and the underlying dorsomedial basal ganglia–thalamus system, selects the representations in the perceptual working memory based on the activations in the executive working memory. We validated the model by showing that it can reproduce and account for a large set of behavioural indices and data related to healthy and pathological participants in the WCST at the state-of-the-art level. These results corroborate and further articulate the proposed hypothesis where the internal manipulation of representations is a core process underlying goal-directed flexible cognition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research received funding from the European Union's Horizon 2020 Research and Innovation Programme under the project 'GOAL-Robots – Goal-based Open-ended Autonomous Learning Robots', Grant Agreement No 713010.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.neunet.2021.07.013.

References

Amos, A. (2000). A computational model of information processing in the frontal cortex and basal ganglia. *Journal of Cognitive Neuroscience*, *12*(3), 505–519.
Aron, A. R. (2007). The neural basis of inhibition in cognitive control. *The Neuroscientist*, *13*(3), 214–228.
Ashby, F. G., Alfonso-Reese, L. A., Waldron, E. M., et al. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review*, *105*(3), 442.
Baldassarre, G., Caligiore, D., & Mannella, F. (2013). The hierarchical organisation of cortical and basal-ganglia systems: a computationally-informed review and integrated hypothesis. In G. Baldassarre, & M. Mirolli (Eds.), *Computational and Robotic Models of the Hierarchical Organisation of Behaviour* (pp. 237–270). Berlin: Springer-Verlag.
Baldassarre, G., Lord, W., Granato, G., & Santucci, V. G. (2019). An embodied agent learning affordances with intrinsic motivations and solving extrinsic tasks with attention and one-step planning. *Frontiers in Neurorobotics*, *13*(45).
Baldassarre, G., Mannella, F., Fiore, V. G., Redgrave, P., Gurney, K., & Mirolli, M. (2013b). Intrinsically motivated action-outcome learning and goal-based action recall: A system-level bio-constrained computational model. *Neural Networks*, *41*, 168–187.
Baldauf, D., & Desimone, R. (2014). Neural mechanisms of object-based attention. *Science*, *344*(6182), 424–427.
Balleine, B. W., & Dickinson, A. (1998). Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. *Neuropharmacology*, *37*(4–5), 407–419.
Barceló, F. (1999). Electrophysiological evidence of two different types of error in the wisconsin card sorting test. *Neuroreport*, *10*(6), 1299–1303.
Barceló, F., & Knight, R. T. (2002). Both random and perseverative errors underlie wcst deficits in prefrontal patients. *Neuropsychologia*, *40*(3), 349–356.
Barceló, F., Suwazono, S., & Knight, R. T. (2000). Prefrontal modulation of visual processing in humans. *Nature Neuroscience*, *3*(4), 399–403.
Barch, D. M., Burgess, G. C., Harms, M. P., Petersen, S. E., Schlaggar, B. L., Corbetta, M., et al. (2013). Function in the human connectome: task-fmri and individual differences in behavior. *Neuroimage*, *80*, 169–189.
Barraclough, D. J., Conroy, M. L., & Lee, D. (2004). Prefrontal cortex and decision making in a mixed-strategy game. *Nature Neuroscience*, *7*(4), 404.
Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology*, *59*, 617–645.
Bengio, Y., Goodfellow, I. J., & Courville, A. (2017). *Deep Learning*. Boston, MA: The MIT Press.
Berdia, S., & Metz, J. (1998). An artificial neural network stimulating performance of normal subjects and schizophrenics on the wisconsin card sorting test. *Artificial Intelligence in Medicine*, *13*(1–2), 123–138.
Berman, K. F., Ostrem, J. L., Randolph, C., Gold, J., Goldberg, T. E., Coppola, R., et al. (1995). Physiological activation of a cortical network during performance of the wisconsin card sorting test: a positron emission tomography study. *Neuropsychologia*, *33*(8), 1027–1046.
Bishara, A. J., Kruschke, J. K., Stout, J. C., Bechara, A., McCabe, D. P., & Busemeyer, J. R. (2010). Sequential learning models for the wisconsin card sort task: Assessing processes in substance dependent individuals. *Journal of Mathematical Psychology*, *54*(1), 5–13.
Bracci, S., Ritchie, J. B., & de Beeck, H. O. (2017). On the partnership between neural representations of object categories and visual features in the ventral visual pathway. *Neuropsychologia*, *105*, 153–164.
Braver, T. S., & Bongiolatti, S. R. (2002). The role of frontopolar cortex in subgoal processing during working memory. *Neuroimage*, *15*(3), 523–536.
Brown, V. J., & Bowman, E. M. (2002). Rodent models of prefrontal cortical function. *Trends in Neurosciences*, *25*, 340–343.
Brunel, N., & Wang, X.-J. (2001). Effects of neuromodulation in a cortical network model of object working memory dominated by recurrent inhibition. *Journal of Computational Neuroscience*, *11*(1), 63–85.
Buschman, T. J., Denovellis, E. L., Diogo, C., Bullock, D., & Miller, E. K. (2012). Synchronous oscillatory neural ensembles for rules in the prefrontal cortex. *Neuron*, *76*(4), 838–846.
Bush, G., Vogt, B. A., Holmes, J., Dale, A. M., Greve, D., Jenike, M. A., et al. (2002). Dorsal anterior cingulate cortex: a role in reward-based decision making. *Proceedings of the National Academy of Sciences*, *99*(1), 523–528.
Caligiore, D., Arbib, M. A., Miall, C. R., & Baldassarre, G. (2019). The super-learning hypothesis: Integrating learning processes across cortex, cerebellum and basal ganglia. *Neuroscience and Biobehavioral Reviews*, *100*, 19–34.
Caligiore, D., Borghi, A., Parisi, D., & Baldassarre, G. (2010). Tropicals: A computational embodied neuroscience model of compatibility effects. *Psychological Review*, *117*(4), 1188–1228.
Carpenter, G. A., & Grossberg, S. (1987). A massively parallel architecture for a self-organizing neural pattern recognition machine. *Computer Vision, Graphics, and Image Processing*, *37*(1), 54–115.

Caso, A., & Cooper, R. P. 2017. A model of cognitive control in the Wisconsin card sorting test: Integrating schema theory and basal ganglia function. In *Proceedings of the 39th Annual meeting of the cognitive science society (CogSci 2017)* London, UK; 16-29 2017, (pp. 210–215).

Caso, A., & Cooper, R. P. (2020). A neurally plausible schema-theoretic approach to modelling cognitive dysfunction and neurophysiological markers in parkinson's disease. *Neuropsychologia*, *140*, Article 107359.

Cella, M., Bishara, A. J., Medin, E., Swan, S., Reeder, C., & Wykes, T. (2014). Identifying cognitive remediation change through computational modelling – Effects on reinforcement learning in schizophrenia. *Schizophrenia Bulletin*, *40*(6), 1422–1432.

Chelazzi, L., Perlato, A., Santandrea, E., & Della Libera, C. (2013). Rewards teach visual selective attention. *Vision Research*, *85*, 58–72.

Chen, B., Xu, T., Zhou, C., Wang, L., Yang, N., Wang, Z., et al. (2015). Individual variability and test-retest reliability revealed by ten repeated resting-state brain scans over one month. *PLoS One*, *10*(12), Article e0144963.

Chersi, F., Mirolli, M., Pezzulo, G., & Baldassarre, G. (2013). A spiking neuron model of the cortico-basal ganglia circuits for goal-directed and habitual action learning. *Neural Networks*, *41*, 212–224.

Cisek, P., & Kalaska, J. F. (2010). Neural mechanisms for interacting with a world full of action choices. *Annual Reviews Neuroscience*, *33*, 269–298.

Clark, A. (1997). *Being There: Putting Brain, Body, and World Together Again*. Boston, MA: MIT press.

Corbetta, M., & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience*, *3*(3), 201–215.

Daselaar, S., Cabeza, R., Ochsne, K., & Kosslyn, S. (2013). Age-related decline in working memory and episodic memory: Contributions of the prefrontal cortex and medial temporal lobes. In *The Oxford Handbook of Cognitive Neuroscience, Vol. 1* (pp. 456–472).

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, *69*(6), 1204–1215.

Dehaene, S., & Changeux, J.-P. (1991). The wisconsin card sorting test: Theoretical analysis and modeling in a neuronal network. *Cerebral Cortex*, *1*(1), 62–79.

Dennis, N., & Cabeza, R. (2012). Frontal lobes and aging: deterioration and compensation. *Principles of Frontal Lobe Function*, *2*, 628–652.

DeYoe, E. A., Carman, G. J., Bandettini, P., Glickman, S., Wieser, J., Cox, R., et al. (1996). Mapping striate and extrastriate visual areas in human cerebral cortex. *Proceedings of the National Academy of Sciences*, *93*(6), 2382–2386.

Diamond, A. (2013). Executive functions. *Annual Review of Psychology*, *64*, 135–168.

Dijkstra, N., Zeidman, P., Ondobaka, S., Gerven, M., & Friston, K. (2017). Distinct top-down and bottom-up brain connectivity during visual perception and imagery. *Scientific Reports*, *7*(1), 5677.

Etkin, A., Egner, T., & Kalisch, R. (2011). Emotional processing in anterior cingulate and medial prefrontal cortex. *Trends in Cognitive Sciences*, *15*(2), 85–93.

Fagg, A. H., & Arbib, M. A. (1998). Modeling parietal-premotor interactions in primate control of grasping. *Neural Networks*, *11*(7–8), 1277–1303.

Fallon, S. J., Mattiesing, R. M., Muhammed, K., Manohar, S., & Husain, M. (2017). Fractionating the neurocognitive mechanisms underlying working memory: independent effects of dopamine and Parkinson's disease. *Cerebral Cortex*, *27*(12), 5727–5738.

Farreny, A., del Rey-Mejías, Á., Escartin, G., Usall, J., Tous, N., Haro, J. M., et al. (2016). Study of positive and negative feedback sensitivity in psychosis using the wisconsin card sorting test. *Comprehensive Psychiatry*, *68*, 119–128.

Felleman, D. J., & Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, *1*(1), 1–47.

Figueroa, I. J., & Youmans, R. J. 2013. Failure to maintain set: a measure of distractibility or cognitive flexibility? In *Proceedings of the human factors and ergonomics society annual meeting, Vol. 57* (pp. 828–832).

Finn, E. S., Shen, X., Scheinost, D., Rosenberg, M. D., Huang, J., Chun, M. M., et al. (2015). Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. *Nature Neuroscience*, *18*(11), 1664–1671.

Fiore, V. G., Sperati, V., Mannella, F., Mirolli, M., Gurney, K., Firston, K., et al. (2014). Keep focussing: striatal dopamine multiple functions resolved in a single mechanism tested in a simulated humanoid robot. *Frontiers in Psychology*, *5*(124), e1–17.

Folstein, J. R., Palmeri, T. J., Van Gulick, A. E., & Gauthier, I. (2015). Category learning stretches neural representations in visual cortex. *Current Directions in Psychological Science*, *24*(1), 17–23.

Frank, M. J., Loughry, B., & O'Reilly, R. C. (2001). Interactions between frontal cortex and basal ganglia in working memory: a computational model. *Cognitive, Affective, & Behavioral Neuroscience*, *1*(2), 137–160.

Fuster, J. M., & Bressler, S. L. (2015). Past makes future: role of pfc in prediction. *Journal of Cognitive Neuroscience*, *27*(4), 639–654.

Gazzaley, A., Clapp, W., Kelley, J., McEvoy, K., Knight, R. T., & D'Esposito, M. (2008). Age-related top-down suppression deficit in the early stages of cortical visual memory processing. *Proceedings of the National Academy of Sciences*, *105*(35), 13122–13126.

Gazzaley, A., & Nobre, A. C. (2012). Top-down modulation: bridging selective attention and working memory. *Trends in Cognitive Sciences*, *16*(2), 129–135.

Gibson, J. J. (1979). *The Ecological Approach to Visual Perception*. Boston, MA: Houghton Mifflin.

Gilbert, S. J., & Shallice, T. (2002). Task switching: A PDP model. *Cognitive Psychology*, *44*(3), 297–337.

Gläscher, J., Adolphs, R., Damasio, H., Bechara, A., Rudrauf, D., Calamia, M., et al. (2012). Lesion mapping of cognitive control and value-based decision making in the prefrontal cortex. *Proceedings of the National Academy of Sciences*, *109*(36), 14681–14686.

Gläscher, J., Daw, N., Dayan, P., & O'Doherty, J. P. (2010). States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron*, *66*(4), 585–595.

Gläscher, J., Hampton, A. N., & O'Doherty, J. P. (2008). Determining a role for ventromedial prefrontal cortex in encoding action-based value signals during reward-related decision making. *Cerebral Cortex*, *19*(2), 483–495.

Goldman-Rakic, P. S. (1996). The prefrontal landscape: implications of functional architecture for understanding human mentation and the central executive. *Philosophical Transactions of the Royal Society B, Biological Sciences*, *351*(1346), 1445–1453.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). Generative adversarial nets. In *Advances in Neural Information Processing Systems* (pp. 2672–2680).

Gottlieb, J. (2007). From thought to action: the parietal cortex as a bridge between perception, action, and cognition. *Neuron*, *53*(1), 9–16.

Grant, D. A., & Berg, E. (1948). A behavioral analysis of degree of reinforcement and ease of shifting to new responses in a weigl-type card-sorting problem. *Journal of Experimental Psychology*, *38*(4), 404.

Gruber, A. J., Dayan, P., Gutkin, B. S., & Solla, S. A. (2006). Dopamine modulation in the basal ganglia locks the gate to working memory. *Journal of Computational Neuroscience*, *20*(2), 153.

Hamby, D. (1994). A review of techniques for parameter sensitivity analysis of environmental models. *Environmental Monitoring and Assessment*, *32*(2), 135–154.

Hartley, A. A., & Speer, N. K. (2000). Locating and fractionating working memory using functional neuroimaging: storage, maintenance, and executive functions. *Microscopy Research and Technique*, *51*(1), 45–53.

Hazy, T. E., Frank, M. J., & O'Reilly, R. C. (2007). Towards an executive without a homunculus: computational models of the prefrontal cortex/basal ganglia system. *Philosophical Transactions of the Royal Society, Series B (Biological Sciences)*, *362*(1485), 1601–1613.

Hearne, L. J., Mattingley, J. B., & Cocchi, L. (2016). Functional brain networks related to individual differences in human intelligence at rest. *Scientific Reports*, *6*(32328).

Heaton, R., Chelune, G., Talley, J., Kay, G., Curtiss, G., di Hardoy, M., et al. (2000). *WCST: Wisconsin Card Sorting Test, Manuale (Forma Completa Revisionata)*. Firenze: Giunti O.S..

Heilbronner, S. R., & Hayden, B. Y. (2016). Dorsal anterior cingulate cortex: a bottom-up view. *Annual Review of Neuroscience*, *39*, 149–170.

Hélie, S., Paul, E. J., & Ashby, F. G. (2012). A neurocomputational account of cognitive deficits in parkinson's disease. *Neuropsychologia*, *50*(9), 2290–2302.

Hinton, G. E. (2012). A practical guide to training restricted boltzmann machines. In *Neural Networks: Tricks of the Trade* (pp. 599–619). Springer.

Hinton, G. E., Osindero, S., & Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, *18*(7), 1527–1554.

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science*, *313*(5786), 504–507.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*, 1735–1780.

Hoffmann, M. (2013). The human frontal lobes and frontal network systems: an evolutionary, clinical, and treatment perspective. *International Scholarly Research Notices, Neurology*, *2013*, Article 892459.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, *79*(8), 2554–2558.

Humphries, M. D., & Prescott, T. J. (2010). The ventral basal ganglia, a selection mechanism at the crossroads of space, strategy, and reward. *Progress in Neurobiology*, *90*(4), 385–417.

Intaitė, M., Noreika, V., Šoliūnas, A., & Falter, C. M. (2013). Interaction of bottom-up and top-down processes in the perception of ambiguous figures. *Vision Research*, *89*, 24–31.

Jeannerod, M., Arbib, M. A., Rizzolatti, G., & Sakata, H. (1995). Grasping objects: the cortical mechanisms of visuomotor transformation. *Trends in Neurosciences*, *18*(7), 314–320.
Kanai, R., & Rees, G. (2011). The structural basis of inter-individual differences in human behaviour and cognition. *Nature Reviews Neuroscience*, *12*(4), 231–242.
Kaplan, G. B., Şengör, N. S., Gürvit, H., Genç, İ, & Güzeliş, C. (2006). A composite neural network model for perseveration and distractibility in the wisconsin card sorting test. *Neural Networks*, *19*(4), 375–387.
Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. Preprint arXiv:1312.6114v10.
Klein, R. M. (2000). Inhibition of return. *Trends in Cognitive Sciences*, *4*(4), 138–147.
Konen, C. S., & Kastner, S. (2008). Two hierarchically organized neural systems for object information in human visual cortex. *Nature Neuroscience*, *11*(2), 224–231.
Kosslyn, S. M. (1999). *Image and Brain (4th Edition)*. Cambridge, MA: The MIT Press.
Kriete, T., Noelle, D. C., Cohen, J. D., & O'Reilly, R. C. (2013). Indirection and symbol-like processing in the prefrontal cortex and basal ganglia. *Proceedings of the National Academy of Sciences*, *110*(41), 16390–16395.
Le Roux, N., & Bengio, Y. (2008). Representational power of restricted Boltzmann machines and deep belief networks. *Neural Computation*, *20*(6), 1631–1649.
Levine, D. S., & Prueitt, P. S. (1989). Modeling some effects of frontal lobe damage – novelty and perseveration. *Neural Networks*, *2*(2), 103–116.
Li, C.-S. R. (2004). Do schizophrenia patients make more perseverative than non-perseverative errors on the wisconsin card sorting test? A meta-analytic study. *Psychiatry Research*, *129*(2), 179–190.
Lie, C.-H., Specht, K., Marshall, J. C., & Fink, G. R. (2006). Using fmri to decompose the neural processes underlying the wisconsin card sorting test. *Neuroimage*, *30*(3), 1038–1049.
Lippmann, R. (1987). An introduction to computing with neural nets. *IEEE Assp Magazine*, *4*(2), 4–22.
Mangun, G. R. (1995). Neural mechanisms of visual selective attention. *Psychophysiology*, *32*(1), 4–18.
Mannella, F., & Baldassarre, G. (2015). Selection of cortical dynamics for motor behaviour by the basal ganglia. *Biological Cybernetics*, *109*, 575–595.
Mannella, F., Gurney, K., & Baldassarre, G. (2013). The nucleus accumbens as a nexus between values and goals in goal-directed behavior: a review and a new hypothesis. *Frontiers in Behavioral Neuroscience*, *7*(135).
Mansouri, F. A., Matsumoto, K., & Tanaka, K. (2006). Prefrontal cell activities related to monkeys' success and failure in adapting to rule changes in a wisconsin card sorting test analog. *Journal of Neuroscience*, *26*(10), 2745–2756.
Mechelli, A., Price, C. J., Friston, K. J., & Ishai, A. (2004). Where bottom-up meets top-down: neuronal interactions during perception and imagery. *Cerebral Cortex*, *14*(11), 1256–1265.
Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, *24*(1), 167–202.
Milner, B. (1963). Effects of different brain lesions on card sorting: The role of the frontal lobes. *Archives of Neurology*, *9*(1), 90–100.
Monchi, O., Petrides, M., Doyon, J., Postuma, R. B., Worsley, K., & Dagher, A. (2004). Neural bases of set-shifting deficits in parkinson's disease. *Journal of Neuroscience*, *24*(3), 702–710.
Monchi, O., Taylor, J. G., & Dagher, A. (2000). A neural model of working memory processes in normal subjects, parkinson's disease and schizophrenia for fMRI design and predictions. *Neural Networks*, *13*(8–9), 953–973.
Naselaris, T., Bassett, D. S., Fletcher, A. K., Kording, K., Kriegeskorte, N., Nienborg, H., et al. (2018). Cognitive computational neuroscience: A new conference for an emerging discipline. *Trends in Cognitive Sciences*, *22*, 365–367.
Nelson, H. E. (1976). A modified card sorting test sensitive to frontal lobe defects. *Cortex*, *12*(4), 313–324.
Nolfi, S., & Floreano, D. (2000). *Evolutionary Robotics: The Biology, Intelligence, and Technology of Self-Organizing Machines*. Cambridge, MA: MIT Press.
Norman, D. (1988). *The Psychology of Everyday Things*. New York: Basic Books.
Nyhus, E., & Barceló, F. (2009). The wisconsin card sorting test and the cognitive assessment of prefrontal executive functions: a critical update. *Brain and Cognition*, *71*(3), 437–451.
Ognibene, D., & Baldassarre, G. (2015). Ecological active vision: four bio-inspired principles to integrate bottom-up and adaptive top-down attention tested with a simple camera-arm robot. *IEEE Transactions on Autonomous Mental Development*, *7*(1), 3–25.
O'Reilly, R. C., & Frank, M. J. (2006). Making working memory work: a computational model of learning in the prefrontal cortex and basal ganglia. *Neural Computation*, *18*(2), 283–328.
O'Reilly, R., & Munakata, Y. (2000). *Computational Explorations in Cognitive Neuroscience – UnderstandIng the Mind By Simulating the Brain*. New York: Bradford book.
Paolo, A. M., Tröster, A. I., Axelrod, B. N., & Koller, W. C. (1995). Construct validity of the wcst in normal elderly and persons with parkinson's disease. *Archives of Clinical Neuropsychology*, *10*(5), 463–473.
Parks, E. L., & Madden, D. J. (2013). Brain connectivity and visual attention. *Brain Connectivity*, *3*(4), 317–338.
Passingham, R. E., & Wise, S. P. (2012). *The Neurobiology of the Prefrontal Cortex: Anatomy, Evolution, and the Origin of Insight, Vol. 50*. Oxford: Oxford University Press.
Perani, D., Schnur, T., Tettamanti, M., Cappa, S. F., Fazio, F., et al. (1999). Word and picture matching: a pet study of semantic category effects. *Neuropsychologia*, *37*(3), 293–306.
Pessoa, L. (2015). Multiple influences of reward on perception and attention. *Visual Cognition*, *23*(1–2), 272–290.
Raffone, A., Srinivasan, N., & van Leeuwen, C. (2014). The interplay of attention and consciousness in visual search, attentional blink and working memory consolidation. *Philosophical Transactions of the Royal Society B, Biological Sciences*, *369*(1641), Article 20130215.
Redgrave, P., Prescott, T. J., & Gurney, K. (1999). The basal ganglia: a vertebrate solution to the selection problem? *Neuroscience*, *89*(4), 1009–1023.
Rentzeperis, I., Nikolaev, A. R., Kiper, D. C., & van Leeuwen, C. (2014). Distributed processing of color and form in the visual cortex. *Frontiers in Psychology*, *5*(932).
Rigotti, M., Ben Dayan Rubin, D. D., Wang, X.-J., & Fusi, S. (2010). Internal representation of task rules by recurrent dynamics: the importance of the diversity of neural responses. *Frontiers in Computational Neuroscience*, *4*(24).
Rizzolatti, G., & Craighero, L. (2004). The mirror-neuron system. *Annual Review of Neuroscience*, *27*, 169–192.
Rizzolatti, G., & Matelli, M. (2003). Two different streams form the dorsal visual system: anatomy and functions. *Experimental Brain Research*, *153*(2), 146–157.
Rougier, N. P., Noelle, D. C., Braver, T. S., Cohen, J. D., & O'Reilly, R. C. (2005). Prefrontal cortex and flexible cognitive control: Rules without symbols. *Proceedings of the National Academy of Sciences*, *102*(20), 7338–7343.
Roy, M., Shohamy, D., & Wager, T. D. (2012). Ventromedial prefrontal-subcortical systems and the generation of affective meaning. *Trends in Cognitive Sciences*, *16*(3), 147–156.
Santucci, V. G., Baldassarre, G., & Mirolli, M. (2016). Grail: A goal-discovering robotic architecture for intrinsically-motivated learning. *IEEE Transactions on Cognitive and Developmental Systems*, *8*(3), 214–231.
Schmidt, R. A. (1976). The schema as a solution to some persistent problems in motor learning theory. In *Motor Control* (pp. 41–65). Elsevier.
Schwarz, G., et al. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.
Seger, C. A. (2008). How do the basal ganglia contribute to categorization? their roles in generalization, response selection, and learning via feedback. *Neuroscience & Biobehavioral Reviews*, *32*(2), 265–278.
Seger, C. A., & Miller, E. K. (2010). Category learning in the brain. *Annual Review of Neuroscience*, *33*, 203–219.
Shankar, S., & Kayser, A. S. (2017). Perceptual and categorical decision making: goal-relevant representation of two domains at different levels of abstraction. *Journal of Neurophysiology*, *117*(6), 2088–2103.
Silvetti, M., Vassena, E., Abrahamse, E., & Verguts, T. (2018). Dorsal anterior cingulate-brainstem ensemble as a reinforcement meta-learner. *PLoS Computational Biology*, *14*, Article e1006370.
Sperati, V., & Baldassarre, G. (2018). A bio-inspired model learning visual goals and attention skills through contingencies and intrinsic motivations. *IEEE Transactions on Cognitive and Developmental Systems*, *10*(2), 326–344.
Srivastava, R. K., Masci, J., Kazerounian, S., Gomez, F., & Schmidhuber, J. (2013). Compete to compute. In *Advances in Neural Information Processing Systems* (pp. 2310–2318).
Steinke, A., Lange, F., & Kopp, B. (2020). Parallel model-based and model-free reinforcement learning for card sorting performance. *Scientific Reports*, *10*(1), 1–18.
Steinke, A., Lange, F., Seer, C., Hendel, M. K., & Kopp, B. (2020b). Computational modeling for neuropsychological assessment of bradyphrenia in parkinson's disease. *Journal of Clinical Medicine*, *9*(4), 1158.
Steinke, A., Lange, F., Seer, C., & Kopp, B. (2018). Toward a computational cognitive neuropsychology of wisconsin card sorts: a showcase study in parkinson's disease. *Computational Brain & Behavior*, *1*(2), 137–150.
Stuss, D., Levine, B., Alexander, M., Hong, J., Palumbo, C., Hamer, L., et al. (2000). Wisconsin card sorting test performance in patients with focal frontal and posterior brain damage: effects of lesion location and test structure on separable cognitive processes. *Neuropsychologia*, *38*(4), 388–402.

Sullivan, E. V., Adalsteinsson, E., Hedehus, M., Ju, C., Moseley, M., Lim, K. O., et al. (2001). Equivalent disruption of regional white matter microstructure in ageing healthy men and women. *Neuroreport*, *12*(1), 99–104.

Sutton, R. S., Barto, A. G., et al. (1998). *Reinforcement Learning: An Introduction*. Boston, MA: MIT press.

Thill, S., Caligiore, D., Borghi, A. M., Ziemke, T., & Baldassarre, G. (2013). Theories and computational models of affordance and mirror systems: An integrative review. *Neuroscience and Biobehavioral Reviews*, *37*, 491–521.

Van Geit, W., De Schutter, E., & Achard, P. (2008). Automated neuron model optimization techniques: a review. *Biological Cybernetics*, *99*, 241–251.

Volpato, C., Schiff, S., Facchini, S., Silvoni, S., Cavinato, M., Piccione, F., et al. (2016). Dopaminergic medication modulates learning from feedback and error-related negativity in parkinson' s disease: a pilot study. *Frontiers in Behavioral Neuroscience*, *10*(205).

Vossel, S., Geng, J. J., & Fink, G. R. (2014). Dorsal and ventral attention systems: distinct neural circuits but collaborative roles. *The Neuroscientist*, *20*(2), 150–159.

Woldorff, M. G., Fox, P., Matzke, M., Lancaster, J., Veeraswamy, S., Zamarripa, F., et al. (1997). Retinotopic organization of early visual spatial attention effects as revealed by pet and erps. *Human Brain Mapping*, *5*(4), 280–286.

Wolters, G., & Raffone, A. (2008). Coherence and recurrency: Maintenance, control and integration in working memory. *Cognitive Processing*, *9*(1), 1–17.

Yin, H. H., & Knowlton, B. J. (2006). The role of the basal ganglia in habit formation. *Nature Reviews Neuroscience*, *7*(6), 464–476.

Yin, H. H., Ostlund, S. B., & Balleine, B. W. (2008). Reward-guided learning beyond dopamine in the nucleus accumbens: the integrative functions of cortico-basal ganglia networks. *European Journal of Neuroscience*, *28*(8), 1437–1448.

Zald, D. H., & Andreotti, C. (2010). Neuropsychological assessment of the orbital and ventromedial prefrontal cortex. *Neuropsychologia*, *48*(12), 3377–3391.

Zanolie, K., Teng, S., Donohue, S. E., van Duijvenvoorde, A. C., Band, G. P., Rombouts, S. A., et al. (2008). Switching between colors and shapes on the basis of positive and negative feedback: An fMRI and EEG study on feedback-based learning. *Cortex*, *44*(5), 537–547.