scientific reports

Check for updates

OPEN

An experimental and computational investigation of executive functions and inner speech in schizophrenia spectrum disorders

Giovanni Granato[1]✉, Raffaele Costanzo[2], Anna Borghi[1,3], Andrea Mattera[1], Sean Carruthers[4], Susan Rossell[4] & Gianluca Baldassarre[1]

Flexible goal-directed human cognition is supported by many forms of self-directed manipulation of representations. Among them, Inner-Speech (IS; covert self-directed speech) acts on second-order representations (e.g., goals/sub-goals), empowering attention and feedback processing. Interestingly, patients with Schizophrenia Spectrum Disorders (SSD) show impaired Executive Functions (EF; e.g., cognitive flexibility) and, probably, a related IS alteration. However, fragmentary evidence and no computational modeling prevent a clear assessment of these processes and focused therapeutic interventions. Here, we address these issues by exploiting a translational approach that integrates experimental clinical data, machine learning, and computational modeling. First, we administered the Wisconsin Cards Sorting Test (WCST; a neuropsychological test probing cognitive flexibility) to 162 SSD patients and 108 healthy control participants, and we computed the clinical behavioural data with a data-driven clustering algorithm. Second, we extracted the cluster neuropsychological profiles with our theory-based validated computational model of the WCST. Finally, we exploited our model to emulate an IS-based psychotherapeutic intervention for SSD subpopulations. We identified different SSD sub-populations and global trends (e.g., a descending feedback sensitivity); however, extremely different neuropsychological profiles emerged. In particular, 'Relatively Intact' patients showed an unexpected profile (distraction/reasoning failures), quite divergent from the perseverative/rigid profile of the others. Importantly, the former showed no impact of Interfering-IS, while the others showed increased Interfering-IS strongly affecting their cognition. These differences highlight that SSD populations require a cluster-dependent individualisation of the intervention to achieve adequate cognitive performance. Overall, these results support a clear definition of neuropsychological profiles and the related Interfering-IS impact in SSD subpopulations, thus showing important implications for basic research (e.g., cognitive neuroscience) and clinical fields (clinical psychology and psychiatry).

Keywords Executive functions, Inner-speech, Schizophrenia spectrum disorders, Computational modeling

Flexible goal-directed cognition is a multi-factorial ability at the basis of human behavioural adaptability. The 'Three-component theory of flexible goal-directed cognition' is a computationally and experimentally validated framework that formalises the neuro-cognitive processes at the basis of cognitive flexibility during goal-directed behaviours[1–3]. It states that flexible cognition depends on the integration between a top-down goal-directed manipulation of representations and sensory-motor interactions with the environment. Specifically, the theory proposes three main features at the basis of flexible goal-directed behaviour. First, three key neuro-functional systems at the basis of the representational manipulation: an executive working memory supported by loops between dorsal cortical regions and ventral basal ganglia[4,5], hierarchical perceptual systems supported

[1]Laboratory of Embodied Natural and Artificial Intelligence, Institute of Cognitive Sciences and Technologies, National Research Council of Italy, Rome, Italy. [2]Department of Movement, Human and Health Sciences, Foro Italico University of Rome, Rome, Italy. [3]Department of Dynamic and Clinical Psychology, and Health Studies, Sapienza University of Rome, Rome, Italy. [4]Faculty of Health, Arts and Design, School of Health Sciences, Centre for Mental Health and Brain Sciences, Swinburne University, Melbourne, Australia. ✉email: giovanni.granato@istc.cnr.it

by perceptual brain hierarchies (e.g., the visual system;[6]), a top-down manipulator supported by frontal-parietal system, basal ganglia[7,8] and language systems[9]. Second, the generation of first-order representations/manipulations (e.g., percepts and selective attention) and second-order representations/manipulations (e.g., goals and inner-speech). Third, the emergence of embodied sensory-motor loops between the agent and the environment. Among many forms of representation manipulation, Inner-Speech (IS; a self-directed covert form of language;[10–12]) targets second-order representations[2], making them more disentangled and shaped. In particular, the representational effect of IS on high-order cognition boosts many cognitive processes such as categorisation[13–15], Executive Functions (EF;[2,11]), working-memory[16], metacognition[17], and motivation[18].

Several studies (for a review see[19]) started investigating the role of IS in psychiatric and neuro-divergent conditions (e.g., Autism;[20]). Here we focus on Schizophrenia Spectrum Disorders (SSDs;[21]), a pool of conditions characterised by 'positive symptoms' (delusions, auditory hallucinations, disorganised speech) and 'negative symptoms' (impaired executive functions and motivation). Among the several deficits in SSD conditions, many studies investigated the relationship between an impaired executive functioning, auditory hallucinations, and IS[19]. For example,[22] correlated auditory hallucinations with an altered IS that might interfere with EF. Moreover,[23] highlighted that IS in SSD patients is fragmented and uncontrolled, thus making higher-order cognitive processes (e.g., cognitive flexibility) more unstable and less effective. On the other hand, some studies[23–25] propose that auditory hallucinations correspond to IS expressions processed by inadequate executive functions (e.g., monitoring), which then lead to a lack of recognition and control of those IS expressions. These scattered findings suggest that impaired IS could lead to different cognitive flexibility alterations in SSD patients (e.g., distraction, thinking fragmentation, motivation alteration, and perseveration). However, the interaction could be the other way around; that is, impaired EF could lead to impaired IS[19].

Overall, experimental evidence[13,19] and our previous computational studies[2,20] corroborate the existence of both a 'supportive IS'—interacting positively with cognitive flexibility, and an 'interfering IS'—worsening cognitive flexibility in different human conditions. Unfortunately, no model proposes and corroborates a unitary vision of this interaction. Consequently, these controversial results prevent a clear understanding and efficient development of related psychotherapies.

In this work, we clarify and formalise the relation between latent components of an impaired flexible cognition in SSDs (e.g., feedback sensitivity, attention, working-memory integrity) and altered IS, potentially stimulating the development of psychotherapeutic interventions. Specifically, we administrated the Wisconsin Cards Sorting Test (WCST; the gold-standard neuropsychological test of executive functions;[26]) to 162 SSD patients and 108 healthy controls, and we computed the clinical behavioural data with a data-driven clustering algorithm (K-means;[27]). Since the data-driven ML method cannot provide insight into the latent causes of behavioural differences (i.e., neuro-cognitive features), we analised the cluster data with our theory-based and experimentally validated computational model of the WCST[1–3,20]. At last, we employed the computational model to emulate the effects of a personalised IS-based psychotherapeutic intervention on patients from different clusters.

Methods

In this section, we present the demographic features of the populations involved in this study and the recruitment/screening procedures. Then, we provide a brief overview of the task setting and scoring. At last, we present the computational techniques we exploited to cluster the clinical data and to profile the participant groups.

Participants and procedures

This study involved an experimental group and a matched control group. All procedures were carried out in strict compliance with the Declaration of Helsinki, participants gave their informed consent to the data re-use, and the ethical commission of the 'School of Health Sciences Centre for Mental Health and Brain Sciences, Swinburne University' approved this study (ref. number 20226934-11953). Participants were recruited from metropolitan-based outpatient and community clinics. Psychiatric diagnosis and HC eligibility were confirmed using the MINI-International Neuropsychiatric Interview[28]. Furthermore, the Wechsler Test of Adult Reading[29] or the National Adult Reading Tests[30] were administered to exclude specific verbal alterations. Participants with significant visual or verbal impairments, a known neurological disorder, and current substance/alcohol abuse or dependence were excluded. At the time of testing, participants were not actively participating in any form of cognitive rehabilitation program or clinical trial.

Participant data were stored in two databases (Cognitive and Genetic Explanations of Mental Illnesses, CAGEMIS; Cooperative Research Center for Mental Health biodatabanks) and all participants had given prior informed consent for the initial analysis of their data. Although we have published a study that includes the original database[31], here we analysed a slightly reduced sample of participants than the original one because some participants of the initial study did not give their informed consent to the data re-use. Therefore, we have excluded 132 participants from the original database. Despite this reduction, the sample size of the control and clinical groups is still large.

Table 1 shows the data of the participants included in this study. The control group is composed of 108 neurotypical participants with an average IQ of 112 (± 10.1) and a mean age of 31 years (± 11.52). The sample shows a balanced sex distribution (54 male vs. 54 female). The experimental group is composed of 162 participants with a diagnosis of SSD, an average IQ of 108 (± 12.5), and a mean age of 38.8 years (± 12.5). The sample shows an imbalanced sex distribution (41 male vs. 121 female), which was also present in the original database.

Task protocol and scoring: the Wisconsin Card Sorting Test

The Wisconsin Card Sorting Test (WCST;[1,26]) is a gold standard neuropsychological test of executive functions, particularly cognitive flexibility. The task requires participants to sort a card deck (128 cards showing a mixed combination of visual items) according to a sorting rule (shape, color, or number; Fig. 1). After each sort,

Group	Sample size	Diagnosis	IQ	Age	Sex
Controls	108	Neurotypical	112 (10.1)	31 (11.52)	F: 54 M: 54
Patients	162	Schizophrenia Spectrum Disorder (DSM-V)	108 (12.5)	38.8 (9.68)	F: 121 M: 41

Table 1. Demographic information on control and experimental groups included in this study. The SSD patients are matched with the control individuals with respect to various features (e.g., IQ, age, etc).

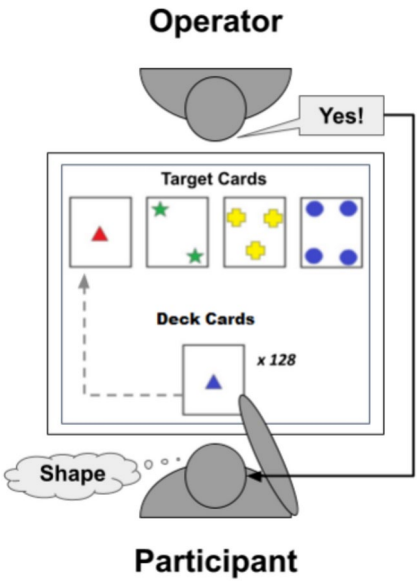

Fig. 1. Task setting and participant-operator interactions. In the example, the participant chooses to sort the deck cards according to the shape rule. In this trial, the shape rule is correct, and the Operator confirms the correctness of this choice. After this feedback, the participant should continue to sort according to the shape rule until the feedback changes.

the operator returns feedback ('Yes' or 'No') to participants. Since the correct sorting rule is never explicitly communicated to participants, they have to infer it only from the operator's feedback. Importantly, the sorting rule changes many times and without warning, so the participants have to infer this change based on the feedback change. Among various behavioural indices proposed by the official documentation, our previous investigations[1,2,20] demonstrated that a specific set of behavioral indices exhaustively describes the participant's cognitive profile. Completed Categories (CC), an index of global performance that identifies the number of times the participant performs a correct change of the sorting rule. Perseverative Errors (PE), that occur when the participant should change the sorting rule and does not do so; it is an index of rigidity/low cognitive flexibility. Non-Perseverative Errors (NPE), all errors that are not scored as PE; it is an index of distraction and reasoning failures. Failures-to-Maintain Set (FMS), all errors that occur after 5 sequential correct responses and before it is necessary to change the rule; it is an index of attentional failures. In this study, we focus on these four indices.

As we did in[1,2,20], we performed multiple statistical comparisons of the task indices (between-models, between-participants, between participants and models). In particular, we ran multiple t-tests and post-hoc analyses (Bonferroni correction) leveraging commonly used python libraries for data analysis and statistics (numpy, pandas, sklearn, scipy).

Clustering and modelling techniques

Here we introduce the clustering methods, the computational model and the procedures we adopted to fit and alter it.

Clusterisation methods

The experimental group's behavioural data underwent a clusterization procedure based on Machine Learning methods (k-means algorithm), extensively explained in[31] and summarised here. This clustering method exploits an unsupervised and agnostic mechanism to discover quantitative similarities between participants' data and categorize them into different groups. Overall, since the algorithm received only behavioural data as input, this procedure does not take into account other individual differences that might influence the clustering operation.

The algorithm is initialised with the WCST indices z-scores of the HC participants and several hierarchical cluster analyses were conducted. This approach ensured the identification and validation of homogeneous performance subgroups within the sample. In particular, collaborative inspection of the agglomeration schedule and dendrogram was used to establish the appropriate number of clusters to be retained and confirmed by discriminant function analysis.

Then, a $k-means$ iterative partitioning (number of iterations = 4) technique was employed to optimise the retained clusters, with initial partitions in the k-means solution defined using the cluster means obtained from the initial clustering procedure.

Last, the stability of the final cluster solution was evaluated through split-sample and alternate method replication via Cohen's analysis, with high agreement over multiple design iterations required to validate the final clustering solution obtained ($K > .80$;[32]).

Computational model

We present here a high-level description of the model neuro-inspired features, components and functioning (Fig. 2); further computational details are reported in Supplementary Materials (a throughout analysis of the algorithms, parameters and their dynamics is presented in[2]). The code of the model is publicly available for online download from the GitHub repository: https://github.com/GiovanniGranato/Flexible-goal-directed-behaviour-and-Inner-Speech.

The model is an instantiation of the 'three-component theory of flexible cognition'[1–3], stating that flexible goal-directed behaviour depends on the integration between a top-down goal-directed manipulation of representations and sensory-motor interactions with the environment. The theory identifies specific neuro-functional systems (hierarchical perceptual systems, executive working memory, top-down manipulator) that underlie representational manipulations and are emulated in the model.

The *hierarchical perceptual component* extracts the input visual features at increasing levels of abstraction (e.g., card features such as colour, shape, and size), emulating the human visual system[6]. This component is implemented with a deep generative model. The *working memory component* stores the 'priorities' of the task sub-goals (e.g., the sorting rules) and determines the probability with which the rules are selected. Working memory priority values undergo a temporal decay, leading them toward the same baseline. In the brain, this function is associated with the reentrant circuits of prefrontal cortices[5]. This component is implemented with a recurrent neural network. The *motivational component* exploits the external feedback to update the information

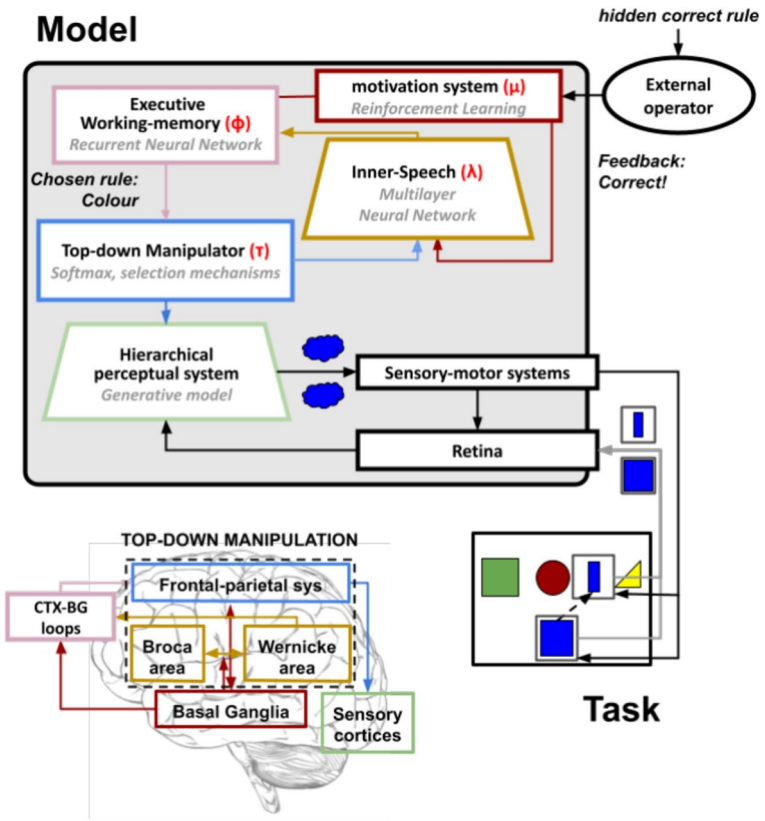

Fig. 2. Architecture of the computational model: algorithms and neuro-inpired features. System-level architecture of the model, emulating various brain networks underpinnings human goal-directed flexible cognition. Key parameters of the model: μ: 'Error sensitivity'; ϕ: 'Memory refresh/Forgetting speed'; τ: 'Explorative tendency/Distractibility'; λ: 'Inner-speech contribution'.

in working memory, emulating the loops between ventral basal ganglia and dorsal prefrontal cortices[33,34]. This component is implemented with a reinforcement learning algorithm. The *selector and manipulator components* support the manipulation of perceptual representations. The former chooses a single sorting rule on the basis of the priorities stored in working memory. The latter implements the manipulation of the internal representations by biasing the perceptual system. The selector and manipulator functions reflect the top-down control that the fronto-parietal cortical system and basal ganglia exert on the perceptual cortices[7,8]. These components are implemented respectively with a softmax function and a disinhibition mechanism. The *IS component* influences the working memory rule selections based on the selector choices and external feedback. This component transmits information regarding the specific working memory rule to be changed (e.g., colour) and the change in valence based on external feedback (positive/negative). The IS component simulates the brain systems that integrate linguistic and emotional information and support human inner-speech[35,36]. This component is implemented as a deep neural network.

The model is also supported by additional components that implement a set of sensorimotor auxiliary functions needed to accomplish the WCST. A *visual sensor component* extracts the visual features from deck cards and target cards. This component emulates the human retina and is implemented with an RGB matrix of pixels. A *visual comparator component* executes a matching of the deck and selected target card on the basis of their manipulated low-level perceptual representations (for a stylised example of this comparison, see Fig. 2). In particular, it computes the Euclidean distance between the rule-based representations of the two cards (e.g., shapeless coloured blobs as abstract representations of colours;[1]). In the brain, these processes rely on the integrated network involving the frontal and temporal-occipital cortices[37,38]. A *motor component* controls the saccades on the deck and target cards and moves the deck card close to the chosen target card after a successful visual matching. This simplified design choice is due to the fact that rule-based category learning is mostly based on a category-based system rather than a procedural motor system[39]. Last, a simulated *'external operator' element* receives the deck card and the chosen target card and returns positive or negative feedback to the agent (this element performs an adequate task scoring, e.g., it knows the trial-by-trial correct sorting rule).

The model has four key free parameters that influence its computations and behaviour. 'Error sensitivity' (μ), representing the magnitude with which the motivational component influences the working-memory priorities in case of negative feedback. 'Memory refresh/Forgetting speed' (ϕ), representing the decay speed of the working memory rule priorities towards a baseline. 'Explorative tendency/Distractibility' (τ), representing the level of randomness of the rule selection. 'Inner-speech contribution' (λ), representing the magnitude with which the IS component influences the working memory. The latter parameter is the most important for this study. Notably, the low number of free parameters is a strong feature of the model.

Development and validation of the model. Several versions of the model are extensively analysed and experimentally corroborated. In[1] we furnished a sensitivity analysis of the first model version (without inner-speech), showing that the parameters were related to different WCST indices. For example, the 'error sensitivity' parameter was mostly negatively correlated to the number of PE. Additionally, parameters "Memory refresh/ forgetting speed' and 'explorative tendency/distractibility' were correlated with the number of NPE. On the other hand, the model reproduced the WCST performance (behavioural indices) of two healthy human populations (young and old adults) and two clinical populations (frontal lesion, Parkinson).

The development, integration and validation of an inner-speech component were supported both by a theory-driven approach (literature evidence) and a data-driven approach (experimental support) in[2]. Many features of the IS component highlight that it adequately mimics the human monological and deliberative/ wilful inner-speech (for an extended analysis of various forms of inner-speech, see[12,40]). First, this component reinforces the system's working memory, emulating the phono-articulatory loop in Baddeley's well-known theory[41]. Second, it emulates the motivational role of inner speech for human cognition. Indeed, different authors have underlined the important motivational and self-reinforcing role of inner speech[42–44]. Third, the function of the inner-speech component is clearly linked to the definition of 'second-order cognition'[45], also called by others 'metacognition'. More recently, metacognition has also been investigated in relation to the strategy changes adopted following error detection[46]. Thus, the metacognitive role of the IS component agrees with a 'second-order form of cognitive control', where one sensorimotor system implicitly represents a feature of another system[47].

On the other hand, the inner-speech component receives a data-driven investigation and validation. First, we performed an extended sensitivity analysis, particularly investigating the effect of the inner-speech component on the WCST indices. Therefore, we discovered that the parameter 'inner-speech contribution' was positively related to global performance indices (e.g., CC) and negatively related to NPE and PE. Second, the model reproduced a complete behavioural profile of three groups of teenagers in different experimental conditions: control, motor tapping, and verbal shadowing (i.e., an experimental protocol disrupting the inner speech contribution). Importantly, the model parameters that best fit the three samples showed a substantially different contribution of inner speech to the task solution (a very lower parameter λ in verbal shadowing condition). Thus, it demonstrated the ability to disentangle the inner-speech contribution during the WCST. Moreover, the results of different lesions to the inner-speech component further highlighted its role in the system (reinforcement of working memory, motivation, and meta-cognition), in line with the theorised role of inner speech. At last, in[20], the model reproduced the behavioural profile of neurotypical and ASC people of different ages. In agreement with the literature proposals[12,19,48–51], the model predicted an increasing age-dependent contribution of inner-speech in neurotypical people and an almost absent, age-independent inner-speech support in Autism Spectrum Condition (ASC) people.

Here, we altered the inner-speech component following a theory-driven approach. In particular, many studies (for a review, see[19]) suggested that the IS could be uncontrolled and altered in SSD, leading to thinking fragmentation, distraction, perseveration, and motivational alterations. Therefore, we produced an IS alteration that shows the key features of an Interfering-IS during the WCST performance and we performed an exhaustive

sensitivity analysis (for computational details of this alteration and investigation, see section "Model-based human group comparisons"). First, the IS alteration randomly affects the trials, thus resulting in a fragmented and uncontrolled effect. Second, it randomly acquires a 'negative feedback effect' in case of external positive feedback. This alteration emulates a distracting effect, thus inducing a shift from the correct response. Third, it randomly acquires a 'positive feedback effect' in case of external negative feedback. This alteration emulates a perseverative effect, thus inducing perseveration toward the incorrect response. Overall, the first feature mimics the uncontrolled effect of an Interfering-IS, while the last two features mimic its motivational alteration.

Taking into account all previous results, different versions of the model have already reproduced human WCST performances in various healthy aging conditions (children, young adults, middle adults, old adults), experimental conditions (control, motor tapping, verbal shadowing), clinical conditions (frontal lesions, Parkinson) and neuro-divergent conditions (autism at different ages). Including the results of the present study, the model reproduced the cognitive and behavioral profiles of 19 human groups.

Fitting procedures, IS alteration and IS-based intervention

We adopt a statistical method to fit the model parameters, namely to find the model parameters that best reproduce the behavioural data of the human groups (for more details, see section 1.2 in Supplementary Materials). In our previous studies[1,2,20], we ran thousands of simulations and we computed the Minimum Square Error (MSE;[52]) to evaluate the fitness of each one. Here, we substituted the MSE with the Bayesian Information Criterion (BIC;[53]). The BIC takes into account the data fit and, importantly, the model complexity (e.g., penalising models with more parameters). Therefore, the BIC is a more suitable index to perform fitting comparisons between different models of the WCST and different versions of a specific model.

Operationalisation of the Interfering-IS and the IS-based psychotherapeutic intervention. We altered the functioning of the neural network at the basis of the IS component, thus producing an Interfering-IS.

First, we randomly changed the feedback provided to this network in two opposite ways: in case of external 'positive feedback', the net receives negative feedback, and vice-versa. Second, we randomised the network activation, in particular setting a 0.3 probability of intervening during a single trial. The Interfering-IS activation exploits the same parameter λ (IS contribution), which in this case identifies the alteration severity (i.e., the interference magnitude with which the Interfering-IS influences the working-memory). We did not change the net input component relating to the selected rule. Overall, according to its theory-based development, the Interfering-IS functioning is characterised by an interference modality, an activation probability and a level of intervention/severity ('Interfering-IS weight').

Starting from the parameter profiles we found during the initial fitting (model with IS), we performed an exhaustive sensitivity analysis involving the new alteration. In particular, we produced many simulations with increasing the 'Interfering-IS weight' and we computed the resulting behavioural effects (WCST indices) and fitting performance (BIC). This approach led to investigate the effect of this new parameter on the model and to find the different Interfering-IS levels that better explain the different clinical sub-group data.

IS-based psychotherapeutic intervention expects a reverse 'alteration', namely a decrease in Interfering-IS weight to 0 and a subsequent increase of the IS parameter. Here, to reproduce the possible effect of the therapy, we assumed that each training session corresponds to a 0.1 decrease/increase of Interfering-IS/IS. Overall, the intervention simulation stops when the clinical model reaches the control model performance (equal/better WCST indices).

Visualisation techniques of the WCST models

Here we optimised an innovative 'visualisation technique' for WCST models, we introduced in[1] and exploited in[2,20].

This technique shows a trail-by-trial complete view of the task administration (see Fig. 12). In particular, it integrates the information about the expected behaviour of the participant (the correct rule he/she should choose), the behavioural response he/she produces (correct response, perseverative errors, etc.), the internal processes of the model (the priorities of the rules that the model stores, which guide its decision-making process). In this work, we insert a further box ('I-IS') that identifies the different Interfering-IS effects.

Overall, this technique offers a useful tool to the scientific and clinical community. In particular, it can support a rapid performance assessment of individual models and across models. In addition, it can facilitate clinicians to use the model as a digital twin of patients.

Results

This section reflects the structure of the work (initial experimental data analysis, clustering output, model validation, model-based extraction of human neuropsychological features, model-based therapeutic predictions) and then proposes multiple data analyses. In particular, we propose primary analyses (behavioural assessment, clustering output, and model-based neuropsychological profiling of groups), secondary analyses (e.g., analyses of individual participant models), and exploratory analyses (analysis of the 'PE-NPE balance' and outcomes of simulated therapeutic interventions).

Experimental data of the control and clinical groups

Patients (n = 162) show statistical differences compared to the control group (N = 108) across all the WCST indices (Table 2).

In particular, the clinical sample shows lower CC (3.4 ± 2.1 VS 5.8 ± 0.6, t = 11.56, Cohen's d = 1.55, $p < 0.001$), higher PE (27.9 ± 18.5 VS 9.0 ± 7.0, t = 10.14, Cohen's d = 1.35, $p < 0.001$), higher NPE (20.5 ± 11.2 VS 8.8 ± 7.2, t = 9.61, Cohen's d = 1.24, $p < 0.001$), and higher FMS (1.6 ± 2.0 VS 0.2 ± 0.5, t = 7.12, Cohen's d = 0.96, $p < 0.001$).

Group	CC	PE	NPE	FMS
Controls (n = 108)	5.8 (0.6)	9.0 (7.0)	8.8 (7.2)	0.2 (0.5)
Patients (n = 162)	3.4 (2.1) ***	27.9 (18.5) ***	20.5 (11.2) ***	1.6 (2.0) ***

Table 2. Behavioural data of control and clinical groups. CC: Completed Categories; PE: Perseverative Errors; NPE: Non-Perseverative Errors; FMS: Failure-to-Maintain Sets.

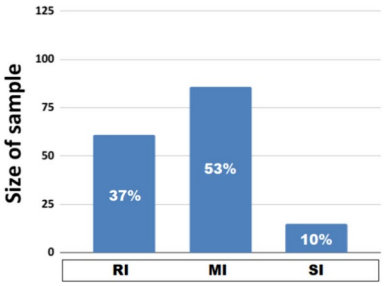

Fig. 3. Sample sizes of the patient sub-groups. RI: Relatively Intact. MI: Moderately Impaired. SI: Severally Impaired. The in-column number refers to the group percentage of the entire clinical group.

Group	CC	PE	NPE	FMS
Relatively Intact (n = 61)	5.84 (0.37)	12.53 (6.23)	11.71 (6.27)	0.98 (1.40)
Moderately Impaired (n = 86)	2.28 (1.04)	35.44 (16.04)	25.15 (9.67)	2.29 (2.28)
Severely Impaired (n = 15)	0.27 (0.59)	47.53 (20.05)	30.07 (11.94)	0.20 (0.78)

Table 3. Patient sub-groups data. Behavioural data of patient sub-groups.

The clustering algorithm found three performance-dependent clinical sub-populations, namely 'Relatively Intact patients' (RI), 'Moderately Impaired patients' (MI), and 'Severely Impaired patients' (SI).

Although the dataset we used here is slightly reduced than the one we used in[31], it kept the same between-group proportions (Fig. 3).

In particular, the RI group (n = 61) represents 37 % of patients, the MI group (n = 86) represents 53 % of them, and the SI group (n = 15) represents 10 % of them. We expected the unbalanced sample sizes and, in particular, the small SI sample size due to (1) the unsupervised clustering algorithm and (2) previous tests on different sample sizes of subpopulations[31]

Clinical sub-groups show a high behavioural variability (Table 3). Indeed, both them and the control group show statistical differences (for a complete list of post-hoc analyses, see Tables S1–S4 in Supplementary Materials). However, some of them show similar indices (Fig. 4).

For example, the healthy control (HC) and RI groups do not show statistically different CC (5.8 ± 0.6 VS 5.84 ± 0.37, t = 0.47, Cohen's d = 0.08, p = 3.83). The MI and SI groups show non-statistically different PE (35.44 ± 16.04 VS 47.53 ± 20.05, t = 2.59, Cohen's d = 0.66, p= .07). The NPE is non-statistically different between the control group and RI (8.8 ± 7.2 VS 11.71 ± 6.27, t = 2.64, Cohen's d = 0.43, p = .05), and between the MI and SI (25.15 ± 9.67 VS 30.07 ± 11.94, t = 1.75, Cohen's d = 0.45, p = .50). Finally, the FMS is non-statistically different between the control group and SI groups (0.2 ± 0.5 VS 0.20 ± 0.78, t = 0.00, Cohen's d = 0.00, p = 6.00) and between RI and SI groups (0.98 ± 1.40 VS 0.20 ± 0.78, t = 2.07, Cohen's d = 0.69, p = 0.25).

We performed a power analysis (with software GPower 3.1.9.7) that took in consideration all behavioural indices and the between-group comparisons (control-clinical groups): HC vs RI (Cohen's d = 0.67, α = 0.05; Power 1–β: 0.99), HC vs MI (Cohen's d = 2.36, α = 0.05; Power 1–β: 1.00), HC vs SI (Cohen's d = 0.8, α = 0.05; Power 1–β: 1.00). All differences observed in the sample show high statistical power (Power > 0.90), supporting the robust results emerging from the cluster analysis.

Modelling experimental data of the control group and patient clusters

The structure of this section aims to investigate the functioning of the model and to validate it. Furthermore, it represents the starting point to offer a model-based analysis of the cognitive profile of the participants. The first section illustrates an initial fitting operation, in which we try to reproduce the experimental data of the participants with the original model (without interfering IS), thus obtaining suboptimal performances for MI

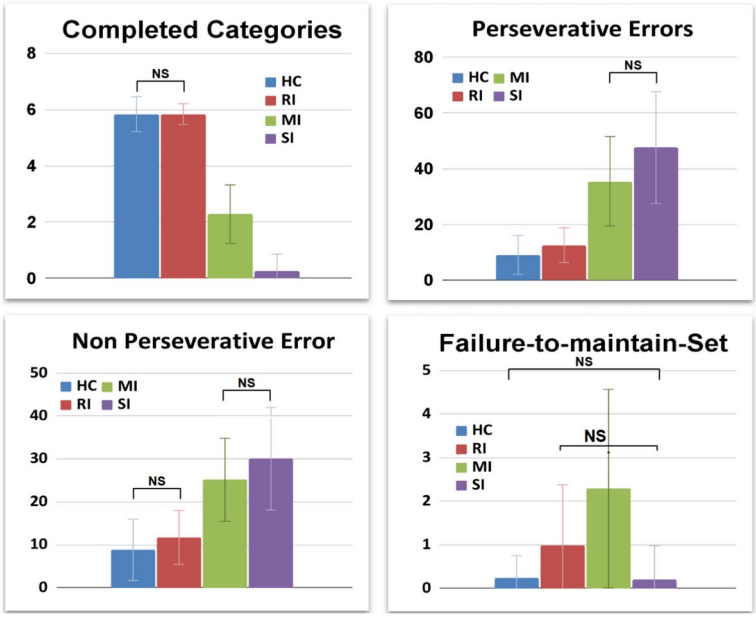

Fig. 4. Behavioural data of healthy/control and patient sub-groups. NS: not significant statistical difference.

and SI populations. The second section illustrates a sensitivity analysis in which we progressively increase the I-IS weight and investigate the subsequent behavioural effect on the model (WCST indices). The last section illustrates a second validation operation, in which we compute the fitting performances (BIC) related to each previous simulation. This last validation achieved impressive performances, leading the model to reproduce even the WCST indices of MI and SI participants with different I-IS weights.

Initial fitting: model-human comparisons

The model reproduces the behavioural trends and behavioural indices of the corresponding human groups (Fig. 5; for the parameter sets of this initial fitting operation see Table S5 in Supplementary Materials).

The control model shows a slight statistical difference for CC (6.00 ± 0.00 VS 5.83 ± 0.62, $t = 2.85$, Cohen's $d = 0.39$, $p = 0.03$, $p < .05$) but no other significant differences in PE (9.01 ± 7.00 VS 8.74 ± 2.32, $t = 0.3805$, Cohen's $d = 0.05$, $p = 4.23$), NPE (8.82 ± 7.15 VS 7.49 ± 3.21, $t = 1.76$, Cohen's $d = 0.24$, $p = .48$), FMS (0.23 ± 0.52 VS 0.38 ± 0.65, $t = 1.87$, Cohen's $d = 0.25$, $p = .38$).

The RI model shows no significant differences in CC (5.84 ± 0.37 VS 5.87 ± 0.42, $t = 0.42$, Cohen's $d = 0.08$, $p = 4.06$), PE (12.53 ± 6.23 VS 12.92 ± 4.47, $t = 0.40$, Cohen's $d = 0.07$, $p = 4.15$), NPE (11.71 ± 6.27 VS 11.79 ± 4.22, $t = 0.08$, Cohen's $d = 0.01$, $p = 5.61$), FMS (0.98 ± 1.40 VS 1.21 ± 1.22, $t = 0.97$, Cohen's $d = 0.18$, $p = 2.01$).

The MI model shows statistical differences in CC (5.19 ± 0.8 VS 2.28 ± 1.04, $t = 20.57$, Cohen's $d = 3.14$, $p < .001$), NPE (15.09 ± 6.68 VS 25.15 ± 9.67, $t = 7.94$, Cohen's $d = 1.21$, $p < .001$) and FMS (0.26 ± 0.49 VS 2.29 ± 2.28, $t = 8.07$, Cohen's $d = 1.23$, $p < .001$). However, it shows no significant difference in PE (35.44 ± 16.04 VS 34.92 ± 8.59, $t = 0.27$, Cohen's $d = 0.04$,$p = 4.75$).

The SI model shows statistical differences in CC (5.27 ± 0.68 VS 0.27 ± 0.59, $t = 21.51$, Cohen's $d = 7.85$, $p < .001$) and NPE (14.73 ± 7.08 VS 30.07 ± 11.94, $t = 4.28$, Cohen's $d = 1.56$, $p = .0012$, $p < .01$). However, it shows no significant differences in PE (47.53 ± 20.05 VS 34.93 ± 5.63, $t = 2.34$, Cohen's $d = 0.86$, $p = 1.58$) and FMS (0.20 ± 0.78 VS 0.53 ± 0.72, $t = 1.20$, Cohen's $d = 0.44$, $p = 1.43$).

In general, the control and RI models reproduce mostly indices and behavioural trends of the corresponding human groups. However, the MI and SI models reproduce only specific trends (PE/NPE balance) and human indices (e.g., PE). Therefore, this first fitting achieves a sub-optimal result (for further details see Table S6 in Supplementary Materials).

Behavioural effects of the IS alteration

The inner-speech alteration has a similar effect on WCST scores across all models, with some model-specific trend differences (Fig. 6). In particular, the alteration leads to a decrease in CC and an increase in PE, NPE and FMS. However, the groups show partially different alteration magnitudes (e.g., see the PE box of Fig. 6) and slopes (e.g., see the FMS box of Fig. 6). Overall, the IS alteration interferes with the baseline performance of the models (see Tables S7–S10 in Supplementary Materials).

Second fitting: reproducing the human data with an Interfering-IS

Different magnitudes of the Interfering-IS lead to models with different fitting performance (Fig. 7).

The Interfering-IS does not improve the fitting performance of the RI model, which already fits all human indices without Interfering-IS. However, the Interfering-IS introduction leads the MI and SI models to fit all

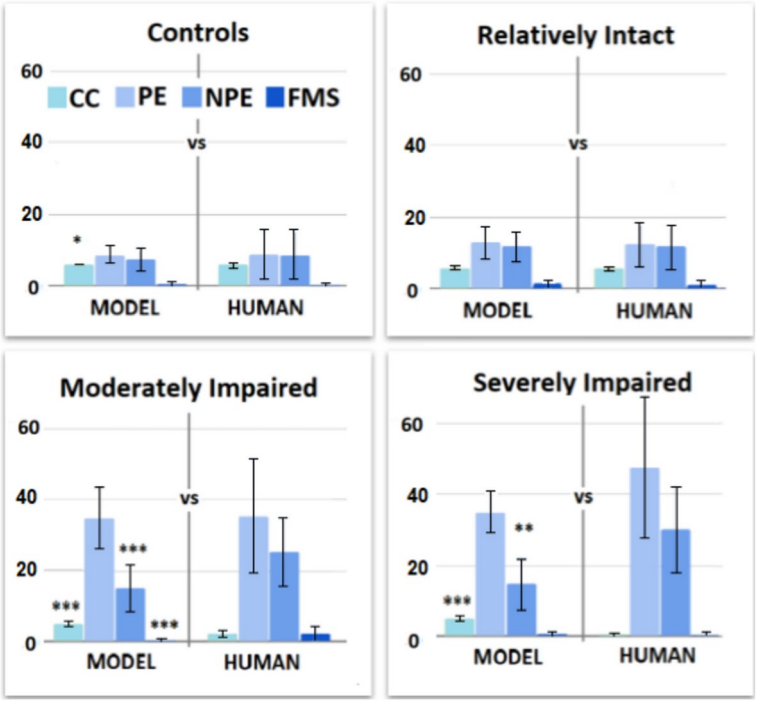

Fig. 5. Behavioural comparisons (Models-Humans). Behavioural profiles of models and corresponding human groups.

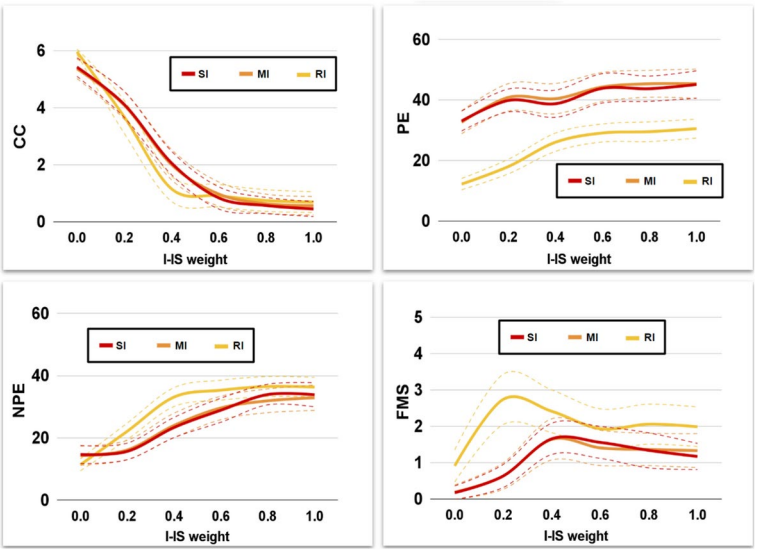

Fig. 6. Behavioural effects of Interfering-IS. Behavioural indices of clinical models after an Interfering-IS injection.

human indices (Fig. 8; for an exhaustive statistical comparison see tables S11–S18 in Supplementary Materials). Indeed, both models achieve a much lower BIC after the alteration.

Model-based human group comparisons

Computational methods support model-based profiling of cognitive and behavioural features of human groups. Therefore, here we show between-model comparisons to highlight between-groups human differences.

Cognitive comparisons

Group-dependent cognitive trends and profiles emerge (Fig. 9).

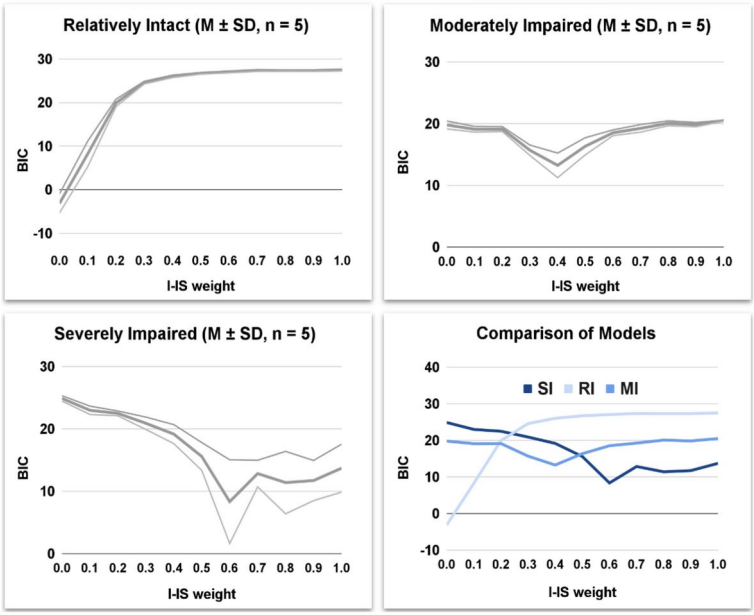

Fig. 7. BIC and Interfering-IS relationship. Fitting performance (BIC) of clinical models with an Interfering-IS injection.

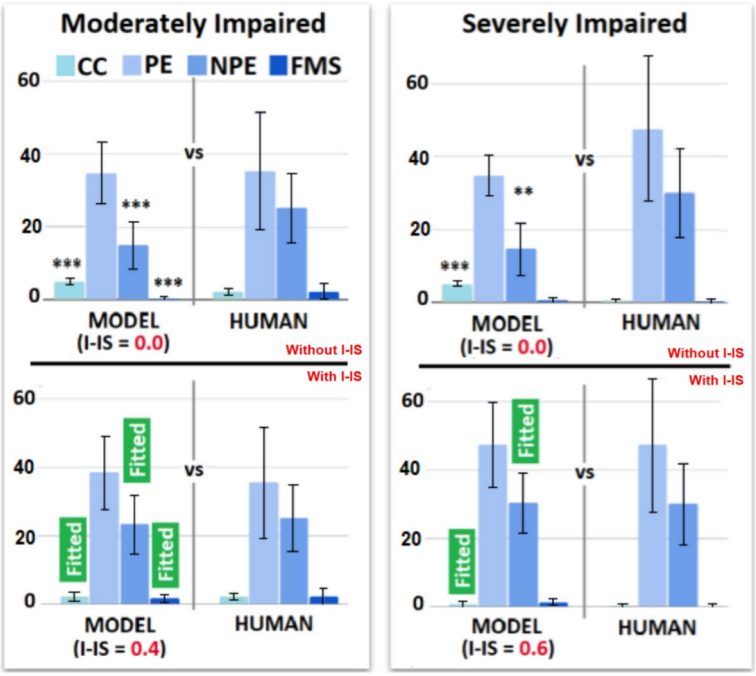

Fig. 8. Behavioural profiles of the IS and Interfering-IS models. The upper plots show the results obtained from the first validation (model without I-IS), which show suboptimal performance (e.g., some WCST indices are not fitted). The lower plots show the results obtained from the second validation (model with I-IS), which show optimal performance. The labels 'Fitted' highlight WCST indices that are fitted in the second validation due to the Interfering-IS introduction.

The 'Error Sensitivity' (μ) shows a severity-based decreasing trend. In particular, the control group shows the highest value (0.27), and clinical groups show gradually lower values (RI: 0.13, MI: 0.08, SI: 0.08). The 'Memory refresh/Forgetting speed' (ϕ) shows an 'irregular severity-based decreasing trend'. The RI group shows the highest value (0.41); however, the control group shows a higher value (0.25) than the MI/SI groups (0.11). The 'Exploratory tendency/Distractibility' (τ) shows a slightly descending trend. The control group shows a higher

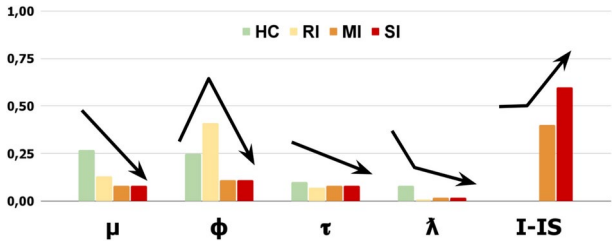

Fig. 9. Parameters of the fitted models. Parameter trends of the models.

Group	CC	PE	NPE	FMS
Controls	6.00 (0.00)	8.27 (2.61)	8.03 (3.05)	0.39 (0.65)
Relatively intact	5.87 (0.42) **	12.92 (4.47) ***	11.79 (4.22) ***	1.21 (1.22) ***
Moderately impaired	2.31 (1.28) ***	38.43 (10.54) ***	23.34 (8.56) ***	1.69 (1.26) ***
Severely impaired	0.73 (0.77) ***	47.47 (12.51) ***	30.27 (8.81) ***	1.27 (0.93) ***

Table 4. Behavioral comparisons between control and clinical models. The asterisks show the result of the t-test between the control model and each of the three clinical models.

value (0.10) compared to the RI group (0.07) and the MI/SI groups (0.08). The 'Inner-Speech contribution' (λ) shows a strongly descending trend. The control group shows the highest value (0.08) compared to mostly null value of the RI group (0.01) and the MI/SI groups (0.02). Notably, the 'Interfering Inner-speech weight' (I-IS) shows a severity-based increasing trend. In particular, the control and RI groups show no IS interference, while the MI group shows a medium value (0.4) and the SI group shows the highest value (0.6).

Overall, the groups show evident severity-based trait trends with some exceptions (e.g., Memory refresh). Interestingly, an increasing Interfering-IS weight specifically differentiates the MI and SI groups.

Behavioural comparisons

The control group shows statistically higher performance compared to clinical groups (Table 4, box A of Fig. 10).

However, macro-similarities emerge between the control and the RI groups and between the MI and the SI groups (box B of Fig. 10; for a complete post-doc analysis see Tables 19–22 in Supplementary Materials).

For example, the control and RI groups show statistical difference in CC (HC: 6.00 ± 0.00 VS 5.87 ± 0.42, t = 3.22, Cohen's d = 0.44, p = .009, p < .01). However, they show very similar values and the statistical difference is due to a zero standard deviation of the HC group. The MI and SI groups show slightly different PE (38.43 ± 10.54 VS 47.47 ± 12.51, t = 2.98, Cohen's d = 0.78, p = .02, p < .05) and NPE (23.34 ± 8.56 VS 30.27 ± 8.81, t = 2.88, Cohen's d = 0.80, p = .03, p < .05), but they are not significantly different in FMS score (1.69 ± 1.26 VS 1.27 ± 0.93, t = 1.23, Cohen's d = 0.38, p = 1.33). Control and SI groups unexpectedly show different FMS score (0.39 ± 0.65 VS 1.27 ± 0.93, t = 4.64, Cohen's d = 1.10, p < .001) than their human counterparts, probably due to a difference-stretching effect performed by the model.

Therefore, although there are clustering-dependent differences, the RI group shows a behavioural profile more similar to the control group. Furthermore, the MI group shows a behavioural profile more similar to the SI group.

PE vs. NPE in control and patient sub-groups. PE and NPE reflect two key opposite neuropsychological trends, that is, perseveration and distraction[1]. For example, their balance could identify specific human neuropsychological profiles (e.g., Autism;[20]). Therefore, we applied these explorative analysis to the participants of this study.

The control and clinical groups show peculiar PE/NPE balances (Fig. 11).

The HC group shows a comparable number of PE and NPE (8.27 ± 2.61 vs 8.03 ± 3.05, t = 0.62, Cohen's d = 0.08, p = 3.21). Similarly, the RI group shows a comparable number of PE and NPE (12.92 ± 4.47 vs 11.79 ± 4.22, t = 1.44, Cohen's d = 0.26, p = .92). The MI group shows significantly higher PE than the NPE (38.43 ± 10.54 vs 23.34 ± 8.56, t = 10.31, Cohen's d = 1.57, p < .001). Similarly, the SI group shows higher PE than the NPE (47.47 ± 12.51 vs 30.27 ± 8.81, t = 4.35, Cohen's d = 1.59, p = .0012, p < .01).

Overall, Fig. 11 (top-left box) shows the differences between PE and NPE (Δ PE-NPE) across the groups (for further details see table S23 in Supplementary Materials). The HC and RI groups show a comparable and almost null value (0.35 ± 3.50 VS 1.13 ± 0.79, t = 1.71, Cohen's d = 0.31, p = .53). The MI and SI groups show a comparable and high value (15.09 ± 13.54 VS 17.2 ± 15.30, t = 0.55, Cohen's d = 0.15, p = 3.52). Therefore,

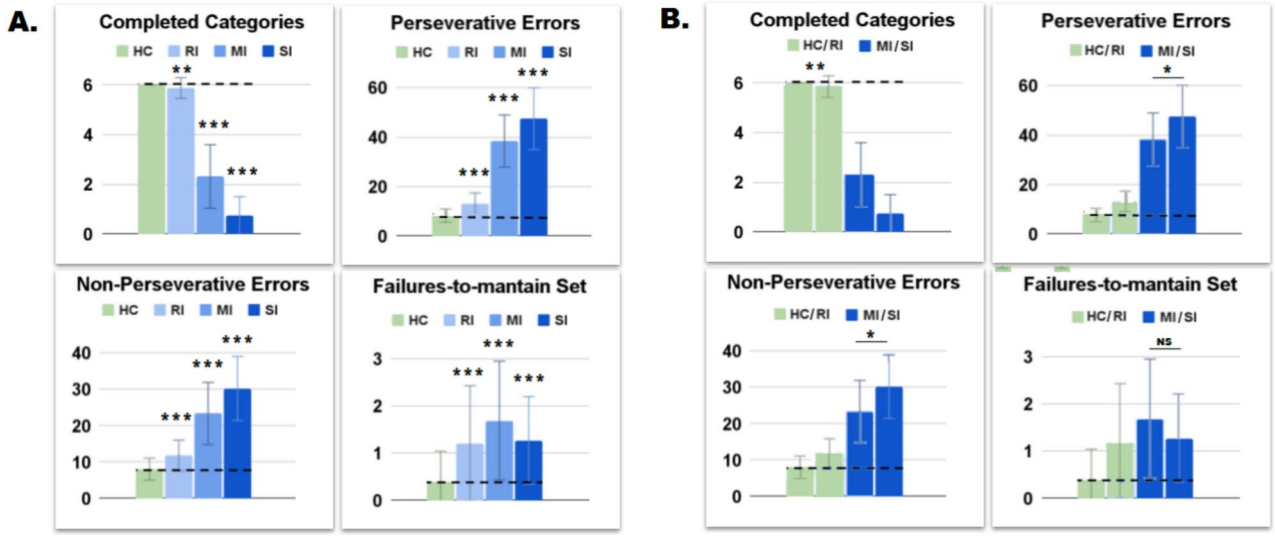

Fig. 10. Behavioural trends of models. A: group-dependent descending behavioral trends. B: macro-similarities in Control-RI models and MI-SI models.

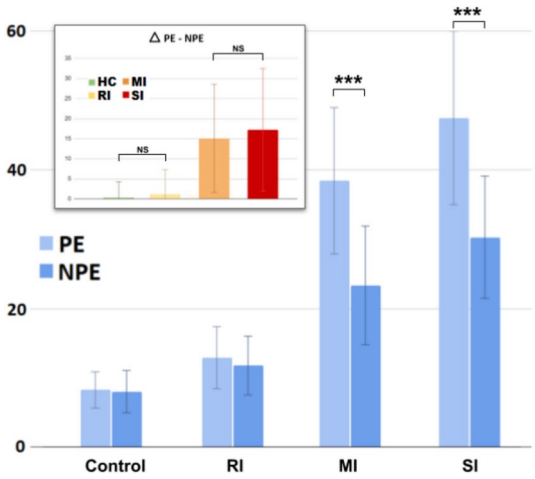

Fig. 11. PE/NPE tendencies. Statistical difference between PE and NPE across the groups. The top-left panel reports the PE-NPE difference Δ.

the HC and RI groups show no PE-NPE imbalance, while the MI and SI groups show the same higher imbalance toward PE.

Internal functioning: single-model analysis

Previous sections have illustrated model-based analyses of the neurocognitive and behavioural features of human groups. Here we analyse the internal processes of the models that reproduce a single participant for each group (Fig. 12).

The HC participant (Fig. 12, top-left plot) adequately performs the task, however, showing some fallacious processes and responses. In particular, it performs adequate reasoning-by-exclusion processes (e.g., trials 0–10 and trials 50–60). However, it shows some attentional failures leading to NPE/FMS (e.g., transient incorrect rule changes; see trials 8 to 15). Moreover, it sometimes shows perseverative tendencies (PE; trials 20–30) and reasoning-by-exclusion process failures leading to both PE and NPE (e.g., trials 73–83). Overall, the participant demonstrates the ability to re-track the correct rule after these failures.

The RI participant (Fig. 12, top-right plot) successfully completes the task, even if showing several distracted/irrational behaviours. In particular, it performs few adequate rule changes (trials 10–20), also showing trends of sustained efficient responses with few errors (trials 70–100 trials). However, it shows many attentional failures (NPE and FMS; trials 20–27 or trials 46–52). Moreover, it often shows severe incorrect rule changes due to impaired reasoning-by-exclusion processes (PE and NPE; trials 34–42 or trials 60–70).

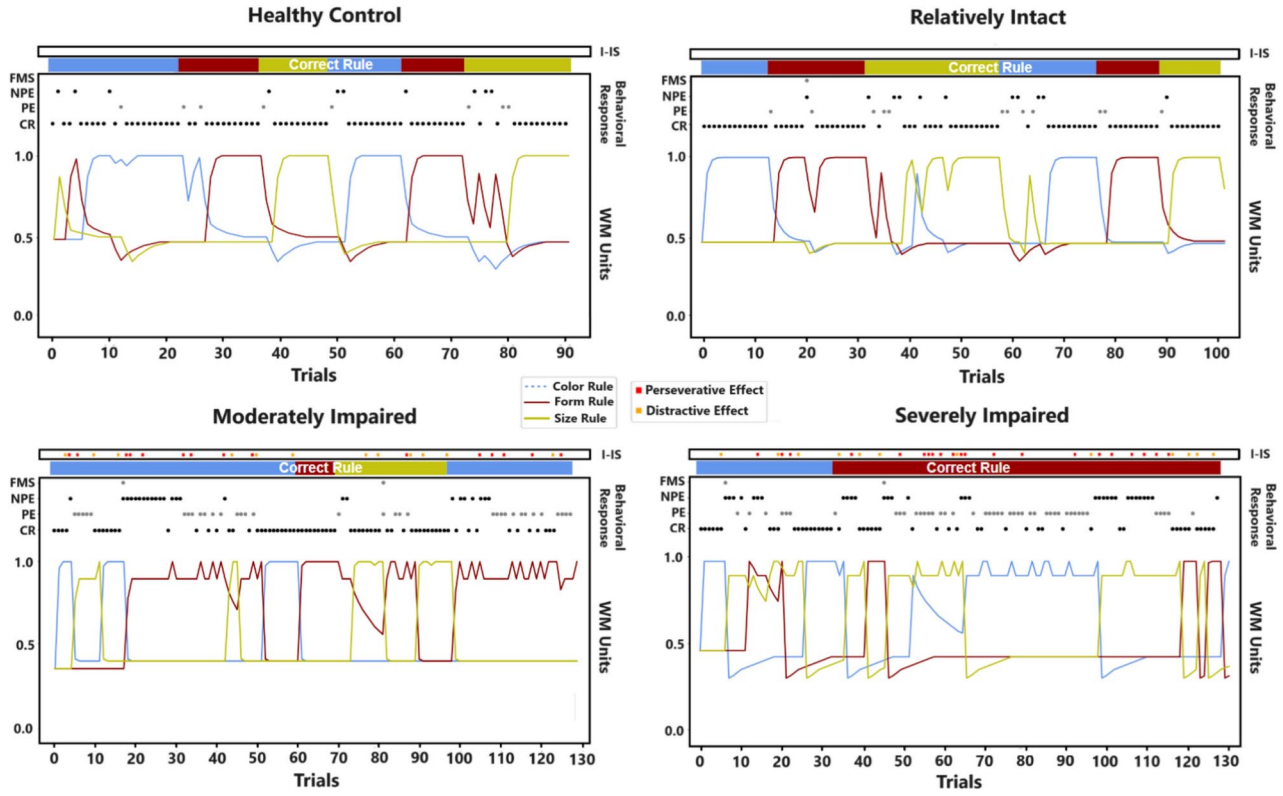

Fig. 12. Internal functioning each of the four single models. The plots show a 'single-participant' view of the models. From the top of each plot: activation/effect of the Inner-Speech component ('I-IS'), the rule that the model should chooses for each trial ('Correct Rule'), the scored behavioural response of the model (FMS, NPE, PE, CR), the internal activation of the working-memory (rule priorities, 0–1).

The MI participant (Fig. 12, bottom-left plot) shows sparse adequate responses (e.g., trials 48–73) but inadequate overall performance; in fact, it fails the task. In particular, it shows attentional failures followed by perseverative responses (trials 1–10 trials or trials 78–88 trials). Moreover, it shows severe perseverative tendencies after incorrect rule changes (trials 18–43), thus leading to both PE and NPE (trials 95–125). Note that the Interfering-IS effect causes attentional failures and subsequent perseverative trends.

The SI participant (Fig. 12, bottom-right plot) mostly shows inadequate responses; indeed, it fails the task. In particular, it shows attentional failures leading to perseverative responses (trials 9–23 or trials 43–72) and severe perseverative trends that also lead to NPE (trials 43–72). Also in case of partially successful rule changes, they result inefficient and lead to NPE (trials 32–39). Note that the Interfering-IS effect mainly causes perseverative trends and also leads to NPE (trials 78–115).

Overall, these results corroborate a group-based descending performance. However, the MI/SI participants show similar deficits (persistent perseverative trends). On the other hand, the RI participant produces a plot similar to the control participant, however showing more attentional failures/distractions rather than perseveration tendencies. At last, these plots highlight that the Interfering-IS causes both perseveration tendencies and attentional failures.

IS-based therapeutic intervention across the groups.

Here we propose exploratory analyses and predictions on a simulated psychotherapeutic intervention (for further data and analyses see tables S24–S28). The simulated intervention has similar improving effects across the clinical groups (Fig. 13, top boxes).

In particular, simulated training sessions (for details, see section 2.3.3) increase the CC, and decrease the PE, NPE, and FMS. However, the groups show different trends and reach control performance at different times (Table 5). The RI group does not show Interfering-IS and requires a minimum potentiation of the IS (0 to 0.1), in particular, only 1 training session to reach the control performance. The MI group requires the recovery of a middle Interfering-IS effect (0.4 to 0) and a higher potentiation of IS (0 to 0.3), thus requiring 7 training sessions to reach the control performance. The SI group requires the recovery of the highest Interfering-IS effect (0.6 to 0) and a high potentiation of the IS (0 to 0.3), thus requiring the most prolonged training intervention (9 training sessions) to reach the control performance.

Overall, the groups show an increasing severity-dependent training time to reach the control performance. Interestingly, the RI model requires little training time to achieve the control performance.

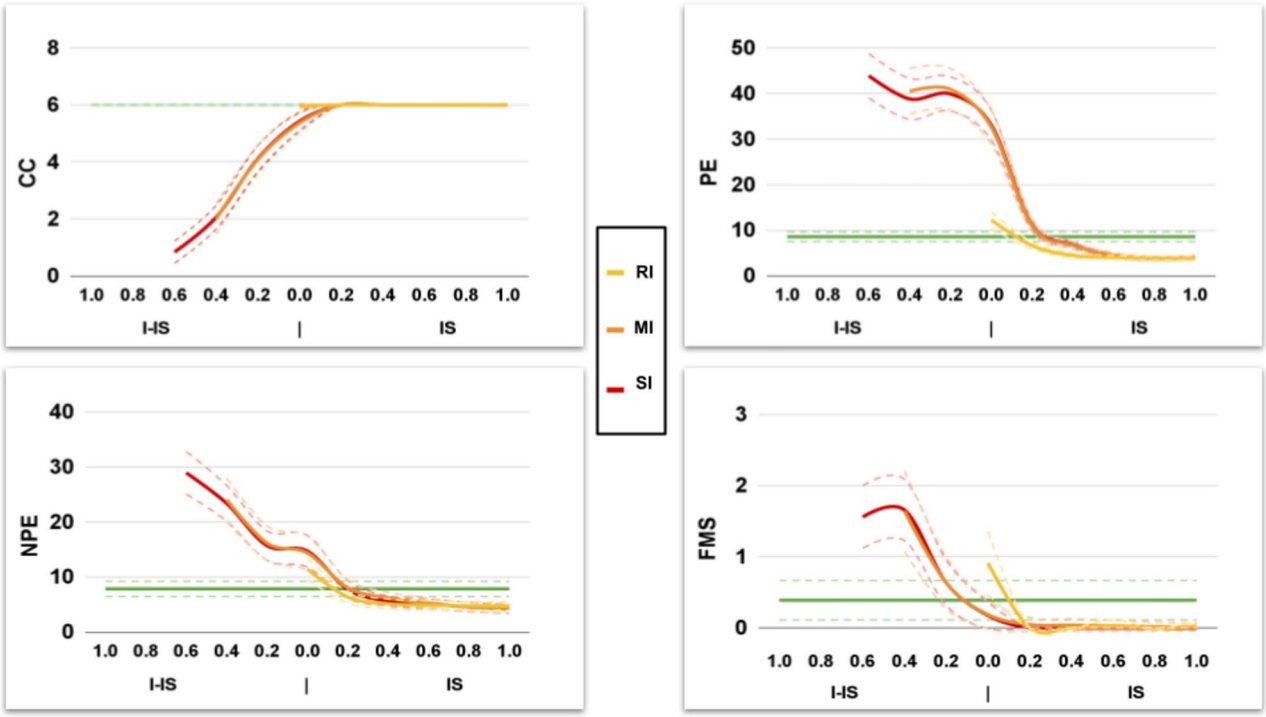

Fig. 13. Effects of an IS-based psychotherapeutic intervention. Trends describing the effect of the intervention based on the patient sub-population. The green line shows the value of the control model.

	Relatively intact	Moderately impaired	Severely impaired
Starting condition	I-IS: 0.0	I-IS: 0.4	I-IS: 0.6
Controls reached at	IS: 0.1	IS: 0.3	IS: 0.3
Training time (sessions)	1	7	9

Table 5. Achievements of a simulated IS-based intervention. Clinical models reach the control performance after different training time (in sessions).

Discussion

SSD patients generally perform worse than HC (Table 2). However, the performed cluster analysis confirmed our previous results[31] as performance-dependent clinical sub-groups emerged (Relatively Intact, RI; Moderately Impaired, MI; Severely Impaired, SI). Importantly, these groups show an irregular distribution (Fig. 3) because most patients fall in the MI group (n = 86, 53 %), and the RI group shows a high sample size (n = 61, 37 %), while few patients fall in the SI group (n = 15, 10 %). These results corroborate the existence of a heterogeneous broad spectrum of schizophrenic conditions[54–56]. However, this group performance does not show a linear trajectory, and between-group similarities emerge (Fig. 4). Indeed, the HC and RI groups show similarities (e.g., global performance), and the MI and SI groups as well (e.g., perseveration). On the other hand, we found a similar FMS value (an index of distraction) between the HC and SI groups. This evidence could highlight an interesting case where the same low FMS value is caused both by healthy performance (control groups typically perform few FMS;[1]) and pathological behavior (an excessive rigidity of the SI group could prevent FMS-related distractions). The imbalance towards PE of SI patients compared to control participants (Fig. 11) supports this hypothesis.

Initial fitting operations validated the computational model and provided initial theoretical insight into these human experimental results. In particular, the initial fitting operation (IS model) reached good performance for the HC and RI groups but reproduced only a few deficits of the MI and SI groups (e.g., perseveration; Fig. 5). However, subsequent emulations demonstrated that the Interfering-IS can get worse the patients' performance (Fig. 6). Therefore, the 'Interfering-IS model' significantly improved fitting performance, thus leading to a successful reproduction of the poor behaviour of the MI and SI groups (Fig. 8). Overall, these results corroborate the idea that an uncontrollable and altered IS (Interfering-IS) is related to altered cognitive flexibility in SSD patients[19,22–24]. In addition, simulations suggest that RI patients do not show an Interfering-IS, while different Interfering-IS weights can distinguish MI and SI participants (Fig. 7). Finally, among different model versions, only the 'Interfering-IS version' successfully reproduces behavioural features of both the control and clinical sub-

groups (see Fig. S1 in Supplementary Materials). Therefore, the Interfering-IS model was demonstrated to be a validated tool for neuropsychological profiling. Indeed, its fitting operations provided a wealth of information on the neuropsychological profile of the participants. The control and clinical groups show a clear descending Error Sensitivity (Fig. 9, first group), thus corroborating the existence of feedback computation deficits in SSD patients[57–60]. A similar descending trend describes the exploratory tendency (Fig. 9, second group); that is, patients show a lower one than the control group. Overall these two trends corroborate the rigidity/perseveration of the SSD patients[61–63]. Importantly, the 'Memory Refresh' shows an irregular trend across the groups (Fig. 9, third group). Indeed, clinical groups show an alteration compared to the control group, thus corroborating a WM impairment in SSD patients[64–66]. However, the MI and SI groups show a lower value (high perseverative tendency) but the RI group shows a higher value compared to the control group. This unexpected result suggests the existence of erratic profiles in SSD patients compared to a common recognise perseverative tendency.

An analysis of the IS contribution (Fig. 9, fourth group) and Interfering-IS weight (Fig. 9, fifth group) of the computational models yields interesting insights regarding the relationship between IS and human cognition.

First, the low IS contribution in the control model corroborates physiological individual differences of the IS contribution across the healthy human population[2,67,67–69]. However, the patient models show almost no IS contribution, suggesting that it does not support the cognitive processes of SSD patients. Secondly, Interfering-IS weight does not show a regular trend between patient models. In fact, the RI group does not show Interfering-IS while the MI and SI groups show a medium/high Interfering-IS. This difference corroborates some variability of the Interfering-IS weight within the human patient population[19]. Furthermore, this result once again indicates a real cognitive difference between the RI and MI/SI human groups. Thirdly, the MI and SI models are differentiated only by the Interfering-IS weight (0.4 and 0.6). This result suggests that (a) MI and SI human groups show a similar neuropsychological profile with different deficit severity and (b) Interfering-IS could be an interfering factor that leads to different and worse patient performance.

Between-models behavioural comparisons corroborated experimental and clustering data: descending behavioural trends emerge (Fig. 10, box A), however HC/RI and MI/SI groups show behavioural similarities (Fig. 10, box B). In particular, behavioural tendencies of the RI and MI/SI look very different. For example, PE-NPE behavioural analyses (Fig. 11) suggest that RI patients have a distraction/perseveration balance. Differently, the MI/SI show a strong imbalance toward perseveration with two increasing severity levels.

An analysis of the single-models internal processes leads to deeper insights into the neuropsychological processes of human participants (Fig. 12). First, as we showed in[1], the HC model can show inefficient reasoning-by-exclusion processes, attentional failures and perseverative tendencies. However, it demonstrates the ability to overcome these difficulties and pass the task. Second, the RI model shows inefficient rule changes (e.g., inefficient reasoning-by-exclusion) rather than perseverative tendencies. Interestingly, as we showed in[1], an 'irrational/distracted' behaviour leads both to NPE and PE. Therefore, many PE can be computed due to reasoning failures rather than perseverative responses (we defined these errors 'distraction-related PE' in[20]). Summing up, the PE/NPE balance of RI models could hide an imbalance toward distraction/reasoning failures, thus supporting more complete interpretations of SSD deficits[70–72]. In contrast, the MI and SI models mostly show extreme perseverative tendencies, which, however, often start with distraction/reasoning errors (e.g., during incorrect rule changes).

The emulation of an IS-based therapeutic intervention confirmed our previous simulation results and led to useful clinical predictions. For example, patient models show group-dependent recovery trajectories to achieve the control model performance (Fig. 13) and require different IS-related interventions (Table 5). In particular, the RI models achieve control performance with short intervention because they are not compromised by an Interfering-IS. On the other hand, the MI and SI models require longer training to reduce the Interfering-IS effect and develop a supportive IS. Therefore, the simulation results predict that an IS-based intervention can lead RI human patients to develop a supportive IS and thus recover their reasoning and attention deficits. On the other hand, supportive IS should lead MI and SI human patients to reduce their perseverative tendencies.

Main contributions: theoretical and computational aspects, clinical relevance and methodological considerations

This work offers many contributions, clinical implications, and methodological considerations. As a main theoretical contribution, it proposes a theory-driven, model-based, and experimentally-validated framework of EF-IS interactions (Fig. 14, box A).

Many studies hypothesised different relationships between an Interfering-IS and altered executive functions[19,22–24]. However, the direction of this interaction and the role of Interfering-IS were unclear. Our proposal supports the idea that EF are intrinsically impaired in SSDs patients (e.g., working memory, inhibitory control, and thus cognitive flexibility; for a review, see[73]). However, an altered interaction between EF (e.g., monitoring) and the IS system generates an Interfering-IS, which in turn worsens EF. Therefore, a vicious loop—rather than a single directional interaction—describes the pathological interaction between EF and IS in SSD patients.

As a main computational contribution, we introduced the first neuro-inspired and experimentally validated computational model that describes the relationship between IS, Interfering-IS, and EF. It represents a research tool to explore further neuropsychological profiles and single cognitive/behavioural traits in populations with a potentially altered IS (e.g., Post-traumatic Stress Disorder and Major Depression Disorder;[74,75]). Indeed, we assess our models with the BIC index—which computes both the fitting performance and the model complexity—to stimulate a community-level comparison between several WCST computational models and populations.

Based on a coherent theoretical framework and a validated computational model, this work offers two main clinical contributions. First, we featured the neuropsychological differences between the RI and MI/SI patients in SSD. Many studies identified an appreciable group of SSD with RI performance, and several hypotheses

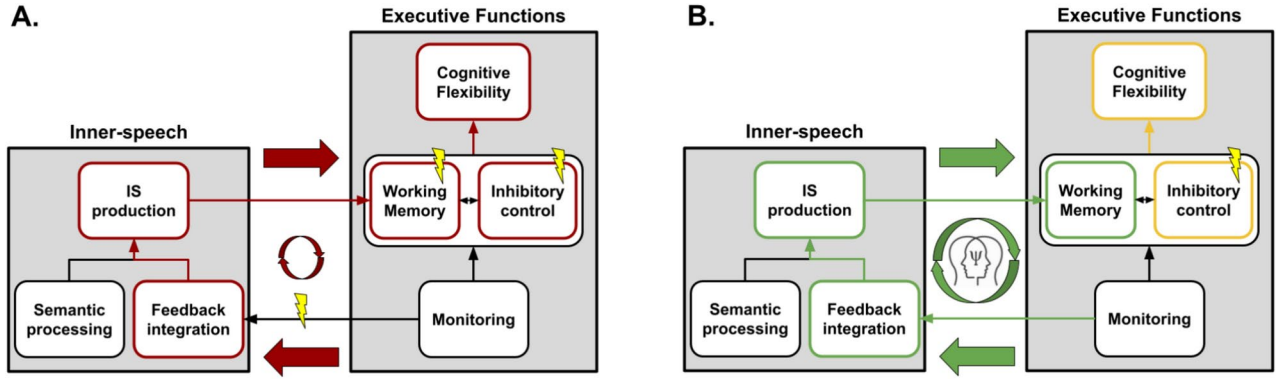

Fig. 14. EF-IS interactions in SSD patients. Yellow lightning bolts represent intrinsic lesions/alterations of SSD patients. Red boxes identify neurocognitive functions that are directly impaired by specific lesions/impairments (e.g., inhibitory control is impaired by an intrinsic deficit) and/or by another identified impaired function (e.g., IS production is impaired by impaired feedback integration). A: pathological EF-IS interactions. B: effect of an IS-based therapeutic treatment.

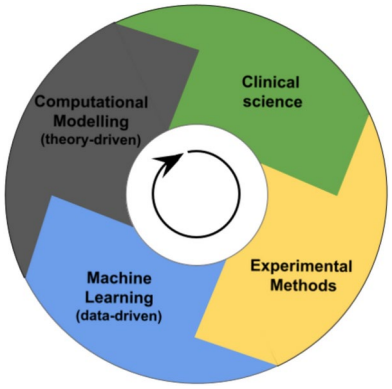

Fig. 15. Translational approach as a key method for basic and applied research. This work exploited a strongly multidisciplinary and translational approach for impacting foundamental and applied research.

emerge[55,76,77]. However, none has given a clear description of their neuropsychological profile. Our results suggest that RI patients show an irrational/distracted profile (attentional failures and inefficient reasoning-by-exclusion) while MI/SI show a perseverative/rigid one. Notably, the RI deficit is limited to the 'rule-change trials'. Second, we emulated an IS-based individualised intervention (Fig. 14, box B). Our findings encourage clinicians to develop psychotherapeutic interventions that directly target IS in SSD and aim to (a) reduce the Interfering-IS onset and the related EF deficit, (b) recover the interaction between an altered self-monitoring competence and IS production, (c) introduce a supportive IS for an intrinsically impaired EF. Despite many psychotherapeutic interventions for SSD patients being proposed (for a review, see[78]), as far as we know, no one directly focuses on IS as a cognitive tool. Our model predicted this intervention might offer individualised beneficial effects for SSD patients.

Finally, this work offers two crucial methodological considerations. First, a multidisciplinary and translational approach is helpful to produce results of interest for both basic and applied research (Fig. 15).

Indeed, our approach integrates clinical science (involvement of patients with schizophrenia spectrum disorders), experimental methods (administration of WCST), machine learning methods (application of a clustering method on WCST data), computational modeling (model-based neuropsychological profiling of neurocognitive features of groups). Importantly, the modeling approach predicts the effects of an IS-based psychotherapeutic intervention, thus bringing benefits to clinical science. In addition, many studies highlight the limitations of purely data-driven machine learning approaches[79,80] and theory-driven approaches in computational psychiatry[81–84]. Here we integrated both a data-driven method (k-means, an unsupervised clustering method) and a theory-driven modelling approach (our WCST model) to counter such limitations and exploit their strengths[85].

Other models of Schizophrenia patients performing the WCST

Many studies developed a computational model that performs the WCST (for a short review, see[1]). However, here we focus on the few that account for Schizophrenia patients and their related results (Table 6).

References	Computational features			Therapy emulation	Data fitted	Free parameters
	System-level architecture	Top-down manipulation	IS/I-IS component			
[86]	✗	✗	✗	✗	✓ (2)	2
[87]	✓	✗	✗	✗	✗	'NA'
[88]	✓/✗	✗	✗	✗	✓ (4)	4
[89]	✗	✗	✗	✗	✓ (3)	3
Granato et al. (2024)	✓	✓	✓	✓	✓ (4)	4

Table 6. WCST computational models of Schizophrenia. 'System-level architecture': emulation approach that emulates several components underlying task performance (e.g., brain structures and cognitive processes). 'Top-down manipulation': first-order/second-order representation manipulation (see the 'three-component theory'). 'IS emulation': emulation of an IS/Interfering-IS component. 'Data fitted': quantitative fitting process and number of fitted human groups. 'Free parameters': hyper-parameters that support the model and the fitting processes. NA: Not Applicable.

[86] proposed a computational model that emulates the synaptic efficiency/stability at the basis of rule selection. The model is supported by a simplified architecture, composed of two mechanisms ('neural noise' and 'gain') modulated by two related parameters, and a rule-selection mechanism. The model validation involved two human populations (healthy and schizophrenic patients) and took into account three behavioural indices (CC, PE, NPE). However, the authors did not performed a quantitative fitting procedure (model-participant comparisons) based on mathematical indices (e.g., MSE). Furthermore, they did not included any therapeutic emulation. Overall, this model is not recent and shows several limitations. For example, it shows excessive architecture simplifications and a limited validation procedure. Therefore, an effective comparison between this model and ours is not feasible. However, we share an interest in the NPE index, which we have demonstrated to be crucial for identifying SSD sub-populations.

[87] proposed a bio-plausible model that reproduces the cortical-subcortical interaction underlying the storage and updating functions of working-memory (e.g., rule selection). The model shows a system-level architecture, but does not provide top-down mechanisms that influence higher-order representations (e.g., inner-speech). The model validation involved a population of schizophrenic patients and focused on only 2 behavioural indices (CC, TE). However, authors did not provided quantitive fitting indices (e.g., MSE) and therapeutic emulations. Although this model shows a system-level architecture, it does not include a top-down mechanism (e.g., Inner Speech) like ours. Furthermore, in contrast to this work, our model is quantitatively validated (BIC) and considers several behavioral indices (CC, PE, NPE, FMS) of four human groups. Based on these fine-grained fitting procedures, in contrast to the model proposed in this work, we predict that both rule storage and selection are compromised in SSD.

[88] proposed a neuro-inspired model that reproduces the dopamine activities underlying cognitive flexibility. The model is composed of three components that simulate the brain macrosystems (frontal cortex, striatum and substantia nigra) and is modulated by four free parameters (frontal gain, frontal bias, striatal gain, striatal output). The model is validated with four human groups (healthy controls and patients with schizophrenia, Parkinson's and Huntington's disease) and two behavioral indices (CC and PE). However, like the two previous models, this work did not exploit a mathematical fitting procedure (e.g., MSE) and did not include any simulation of therapeutic intervention. While this model emulates specific interacting brain systems, ours exhibits a system-level architecture (e.g., includes perceptual and motor components) and emulates additional key interactions (e.g., perception and IS). On the other hand, this model and ours report different predictions. For example, this model predicts that PEs are related to cortical dysfunction while NPEs are related to striatal dysfunction. In contrast, our model predicts that cortical and subcortical dysfunctions can lead to both PEs and NPEs, and that counterintuitive phenomena emerge (e.g., PE related to distraction;[20]).

[90] proposed a computational model that reproduces the dopamine-based RL processes underlying cognitive flexibility. The model shows a simplified architecture that emulates cortical/subcortical processes and is modulated by three parameters (reward sensitivity, punishment sensitivity, and choice consistency).[89] applied this model to reproduce the behaviour of three human populations: healthy participants and two groups of schizophrenic participants (with and without cognitive therapy). The model's fit performance was assessed with log-likelihood estimation and three indices (CC, TE, PE) of the populations were reproduced. This study infers RL deficits in SSD patients and beneficial effects of cognitive therapy. In addition to several architectural oversimplifications of this model, our model shows some strengths compared to it: it also considers the NPE index (a key index to infer reasoning/attention deficits), takes into account SSD subprofiles, and predicts possible psychotherapeutic effects.

Compared to these models, ours shows a more complex architecture and fine-grained fitting methods, which led to a more comprehensive quantitative/qualitative analysis of health conditions and sub-profiles of schizophrenia. Indeed, we quantitatively fitted four human groups with four free parameters. Differently from the models mentioned above, our model fit was assessed with more advanced methods (e.g. BIC approach), evaluating the complexity of the model (number of free parameters) together with the quality of data fitting. Finally, we emulated a psychotherapeutic intervention.

Limitations and future directions

Section "Main contributions: theoretical and computational aspects, clinical relevance and methodologicalconsiderations" shows that this work offers several contributions to basic and applied research. In addition, section "Other models of Schizophrenia patients performing the WCST" shows that our WCST model is the only one that emulates EF-IS interactions and reproduces a complete behavioural profile of four human groups (CC, PE, NPE, FMS). The model thus extends the state-of-the-art in relevant directions. However, this work presents potential limitations we plan to overcome, thus turning them into development directions.

Theoretical/computational aspects: IS, Interfering-IS and their dynamics

The Interfering-IS component has followed a theory-based/evidence-based development; thus it shows the key commonly recognised features of an Interfering-IS (fragmented and uncontrolled onset, motivational alterations, individual differences; for extended reviews see[19,67]). In particular, we fixed an onset frequency and the motivational alteration, thus we altered the Interfering-IS weight to fit the patient data. This simplified approach allowed a flexible integration of the Interfering-IS in our model, thus introducing a clear description of Interfering-IS and EF interactions. However, in the future we could dissociate and parameterise the two motivational alterations (distracting and perseverative effects) and the onset frequency. Although computationally demanding, this dissociation could allow the investigation of different Interfering-IS forms across the clinical population.

The initial fitting operation (IS model) suggested that SSD patients are not supported by a beneficial IS. They may not use their IS due to an insight on its corruption (e.g., insight on their auditory hallucination;[91]) or due to functional alterations of the brain structures at the basis of the supportive IS[92]. Therefore, we replaced the IS component with an Interfering-IS component and performed parameters research and sensitivity analysis with different Interfering-IS levels. The Interfering-IS model showed the best-fitting performance (reproducing all behavioural indices of the three human clinical groups) and corroborated a key idea (supportive IS is not functional in SSD patients), thus justifying our IS replacement with an Interfering-IS. However, we could hypothesise that some patients contemporarily show a residual IS and an Interfering-IS. Although very computationally intensive, in the future, we could consider a massive parameter search that includes multiple contextual parameter alterations (e.g., IS parameter, Interfering-IS parameter, and other parameters). This operation may support the investigation of further hypotheses about the neurocognition of SSD patients.

Methodological aspects: task latent features and participants

In[2,20], we demonstrated that the four WCST indices we consider (CC, PE, NPE, FMS) support an efficient and complete neuropsychological analysis of the participants. However, in the future additional behavioural indices may be considered to extract further participant traits (e.g., the learning-to-learning index;[26]). Anyway, modeling methods require a cost-benefit trade-off, so we should focus on the minimum number of indices that return the maximum amount of information on participants.

The sample size of the control group (n = 108) and SSD entire group (n = 162) is large. However, SSD subgroups show a non-large sample size (e.g., the SI patient subgroup shows N = 15). We should collect more SSD patients, thus achieving a higher sample size, especially for SI patients. However, as emerged from our original investigations[31], SI patients are rarer in the SSD general population[54–56] and will always show a smaller sample size compared to the others. On the other hand, the SSD group shows a sex-related imbalance as in our original dataset[31]. The existence of sex-related differences in executive functions in SSD patients is still debated[93], however in our previous study female participants showed partially better performance (e.g., high CC). Although an investigation of sex/age-specific effects is beyond the scope of this study and would have required too high computational costs, we will leverage our computational methods to perform them on clinical samples in the future (e.g., age/sex focused clustering algorithm initialization and subsequent modeling).

Behavioural data are collected while administering a WCST standard form[26]. This form requires the cooperation of many cognitive processes and prolonged sustained attention, thus leading to a high cognitive load compared to other forms (e.g.,[94,95]). However, this version does not include a condition/method to assess the inner-speech contribution directly. Therefore, as in[20], the model predicts the IS/Interfering-IS onset on the basis of the participant behavioural data. These predictions should be tested with an adequate experimental protocol. Importantly,[68] proposed a WCST protocol that includes three different conditions (control, motor tapping and verbal shadowing), thus directly evaluating the inner-speech contribution to the task performance. Our model reproduced the behavioral data collected during each of the three conditions, modeling the underlying cognitive processes[2]. In particular, it demonstrated disentangling the inner-speech contribution during the control and verbal shadowing conditions. In the future, we may consider the administration of the 'WCST + verbal shadowing' to SSD patients. However, this protocol is particularly effortful for patients with reasoning/attention deficits; thus, feasibility studies should be performed in advance.

Further future directions: a research agenda

In addition to the discussed developments, we aim to develop these investigations further to increase their impact on the scientific and clinical communities.

For example, we aim to develop collaborations with experimentalists and psychotherapists, thus supporting translational projects. In particular, our modeling approach can support model-based analysis of cognitive training interventions and classical psychotherapeutic approaches developed by clinicians. Overall, this direction is supported by a translational approach to basic and applied research.

On the other hand, many research directions would stimulate a high impact of these investigations on the scientific community. For example, cross-patients computational studies would disentangle the IS role in several clinical and neuro-divergent conditions (Autism, Post-traumatic Stress Disorder, Major Depression Disorder,

etc). Integrating these studies with targeted investigations of group-based differences (e.g., sex and age effects) will support a more precise model-based neuropsychological assessment, with positive clinical implications.

Furthermore, a community-level comparison between several WCST models and human groups would lead to the collection of more consistent evidence. Our exploitation of the BIC index points in this direction. Additionally, a shared web-based summary of WCST-related modeling studies could stimulate researchers to compare their findings with the community.

Last, a validated model and shared datasets would lead to the development of a user-friendly research tool for the WCST data analysis. In this respect, we are integrating our model in the online research platforms EBRAINS[96,97], thus boosting the WCST scoring efficacy and its exploitation in research and clinical settings.

Data availability

All data generated or analysed during this study are included in this published article and its supplementary information files.

Received: 17 June 2024; Accepted: 6 February 2025
Published online: 12 February 2025

References

1. Granato, G. & Baldassarre, G. Internal manipulation of perceptual representations in human flexible cognition: a computational model. *Neural Netw.* **143**, 572–594 (2021).
2. Granato, G., Borghi, A. M. & Baldassarre, G. A computational model of language functions in flexible goal-directed behaviour. *Sci. Rep.* **10**, 21623 (2020).
3. Granato, G. & Baldassarre, G. Bridging flexible goal-directed cognition and consciousness: the goal-aligning representation internal manipulation theory. *Neural Netw.* **176**, 106292 (2024).
4. Gruber, A. J., Dayan, P., Gutkin, B. S. & Solla, S. A. Dopamine modulation in the basal ganglia locks the gate to working memory. *J. Comput. Neurosci.* **20**, 153 (2006).
5. Barraclough, D. J., Conroy, M. L. & Lee, D. Prefrontal cortex and decision making in a mixed-strategy game. *Nat. Neurosci.* **7**, 404 (2004).
6. Konen, C. S. & Kastner, S. Two hierarchically organized neural systems for object information in human visual cortex. *Nat. Neurosci.* **11**, 224–231. https://doi.org/10.1038/nn2036 (2008).
7. Redgrave, P., Prescott, T. J. & Gurney, K. The basal ganglia: a vertebrate solution to the selection problem?. *Neuroscience* **89**, 1009–1023 (1999).
8. Gazzaley, A. & Nobre, A. C. Top-down modulation: bridging selective attention and working memory. *Trends Cogn. Sci.* **16**, 129–135 (2012).
9. Geva, S. et al. The neural correlates of inner speech defined by voxel-based lesion-symptom mapping. *Brain* **134**, 3071–3082 (2011).
10. Borghi, A. M. *The Freedom of Words: Abstractness and the Power of Language* (Cambridge University Press, 2023).
11. Langland-Hassan, P. & Vicente, A. *Inner Speech: New Voices* (Oxford University Press, 2018).
12. Alderson-Day, B. & Fernyhough, C. Inner speech: development, cognitive functions, phenomenology, and neurobiology. *Psychol. Bull.* **141**, 931 (2015).
13. Borghi, A. M. & Fernyhough, C. Concepts, abstractness and inner speech. *Philos. Trans. R. Soc. B* **378**, 20210371 (2023).
14. Foerster, F. R., Borghi, A. M. & Goslin, J. Labels strengthen motor learning of new tools. *Cortex* **129**, 1–10 (2020).
15. Boutonnet, B. & Lupyan, G. Words jump-start vision: a label advantage in object recognition. *J. Neurosci.* **35**, 9329–9335 (2015).
16. Baddeley, A. Working memory. *Science* **255**, 556–559 (1992).
17. Clark, A. Magic words: How language augments human computation. In *Language and Meaning in Cognitive Science* 21–39 (Routledge, 2012).
18. Morin, A. The self-reflective functions of inner speech: thirteen years later. In *Inner Speech: New Voices* 276–298 (2018).
19. Petrolini, V., Jorba, M. & Vicente, A. The role of inner speech in executive functioning tasks: schizophrenia with auditory verbal hallucinations and autistic spectrum conditions as case studies. *Front. Psychol.* **2020**, 2452 (2020).
20. Granato, G., Borghi, A. M., Mattera, A. & Baldassarre, G. A computational model of inner speech supporting flexible goal-directed behaviour in autism. *Sci. Rep.* **12**, 14198 (2022).
21. Association, A. P. *Diagnostic and Statistical Manual of Mental Disorders: DSM-5* (American Psychiatric Association Arlington, 2013).
22. Alderson-Day, B. et al. Shot through with voices: dissociation mediates the relationship between varieties of inner speech and auditory hallucination proneness. *Conscious. Cogn.* **27**, 288–296 (2014).
23. Langland-Hassan, P. Fractured phenomenologies: thought insertion, inner speech, and the puzzle of extraneity. *Mind Lang.* **23**, 369–401 (2008).
24. Vicente, A. The comparator account on thought insertion, alien voices and inner speech: some open questions. *Phenomenol. Cogn. Sci.* **13**, 335–353 (2014).
25. Frith, C. D. *The Cognitive Neuropsychology of Schizophrenia* (Psychology press, 2015).
26. Heaton, R. K. & Staff, P. Wisconsin card sorting test: computer version 2. *Odessa: Psychol. Assess. Resourc.* **4**, 1–4 (1993).
27. Ahmed, M., Seraj, R. & Islam, S. M. S. The k-means algorithm: a comprehensive survey and performance evaluation. *Electronics* **9**, 1295 (2020).
28. Sheehan, D. V. et al. The mini-international neuropsychiatric interview (mini): the development and validation of a structured diagnostic psychiatric interview for dsm-iv and icd-10. *J. Clin. Psychiatry* **59**, 22–33 (1998).
29. Wechsler, D. *Wechsler Test of Adult Reading: WTAR* (Psychological Corporation, 2001).
30. Blair, J. R. & Spreen, O. Predicting premorbid iq: a revision of the national adult reading test. *Clin. Neuropsychol.* **3**, 129–136 (1989).
31. Carruthers, S. P. et al. Exploring heterogeneity on the wisconsin card sorting test in schizophrenia spectrum disorders: a cluster analytical investigation. *J. Int. Neuropsychol. Soc.* **25**, 750–760 (2019).
32. Landis, J. R. & Koch, G. G. The measurement of observer agreement for categorical data. *Biometrics* **1977**, 159–174 (1977).
33. Gläscher, J., Daw, N., Dayan, P. & O'Doherty, J. P. States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron* **66**, 585–595 (2010).
34. Mannella, F., Gurney, K. & Baldassarre, G. The nucleus accumbens as a nexus between values and goals in goal-directed behavior: a review and a new hypothesis. *Front. Behav. Neurosci.* **7**, 135 (2013).
35. Kotz, S. A., Meyer, M. & Paulmann, S. Lateralization of emotional prosody in the brain: an overview and synopsis on the impact of study design. *Prog. Brain Res.* **156**, 285–294 (2006).
36. Sidtis, J. J., Van-Lancker-Sidtis, D., Dhawan, V. & Eidelberg, D. Switching language modes: complementary brain patterns for formulaic and propositional language. *Brain Connect.* **8**, 189–196 (2018).

37. Perani, D. et al. Word and picture matching: a pet study of semantic category effects. *Neuropsychologia* **37**, 293–306 (1999).
38. Kosslyn, S. M. *Image and Brain: The Resolution of the Imagery Debate* (MIT press, USA, 1996).
39. Ashby, F. G. & Valentin, V. V. Multiple systems of perceptual category learning: Theory and cognitive tests. In *Handbook of Categorization in Cognitive Science* 157–188 (Elsevier, 2017).
40. Perrone-Bertolotti, M., Rapin, L., Lachaux, J.-P., Baciu, M. & Loevenbruck, H. What is that little voice inside my head? inner speech phenomenology, its role in cognitive performance, and its relation to self-monitoring. *Behav. Brain Res.* **261**, 220–239 (2014).
41. Baddeley, A. Working memory: an overview. *Working Memory Educ.* **2006**, 1–31 (2006).
42. Morin, A. *Inner Speech: New Voices, The Self-Reflective Functions of Inner Speech: Thirteen Years Later* 276–298 (Oxford University Press, 2018).
43. Hardy, J., Hall, C. R. & Hardy, L. Quantifying athlete self-talk. *J. Sports Sci.* **23**, 905–917 (2005).
44. McCarthy-Jones, S. & Fernyhough, C. The varieties of inner speech: Links between quality of inner speech and psychopathological variables in a sample of young adults. *Conscious. Cogn.* **20**, 1586–1593 (2011).
45. Clark, A. Magic words: How language augments human computation. *Lang. Thought: Interdiscipl. Themes* **1998**, 162–183 (1998).
46. Yeung, N. & Summerfield, C. Metacognition in human decision-making: confidence and error monitoring. *Philos. Trans. R. Soc. B: Biol. Sci.* **367**, 1310–1321 (2012).
47. Borghi, A. M., Fini, C. & Tummolini, L. Abstract concepts and metacognition: searching for meaning in self and others. In *Handbook of Embodied Psychology* 197–220 (Springer, 2021).
48. Kray, J., Eber, J. & Lindenberger, U. Age differences in executive functioning across the lifespan: the role of verbalization in task preparation. *Acta Physiol. (Oxf.)* **115**, 143–165 (2004).
49. Fry, P. S. Assessment of private and inner speech of older adults in relation to depression. In *Private Speech: From Social Interaction to Self-Regulation* 267–284 (1992).
50. John-Steiner, V. Private speech among adults. In *Private Speech* 295–306 (Psychology Press, 2014).
51. Williams, D. M., Peng, C. & Wallace, G. L. Verbal thinking and inner speech use in autism spectrum disorder. *Neuropsychol. Rev.* **26**, 394–419 (2016).
52. Johnson, D. Minimum mean squared error estimators. *Connexions* (2004).
53. Millar, P. *et al.* Using the bayesian information criterion (bic) to judge models and statistical significance. In *North American Stata Users' Group Meetings 2006* 1 (Stata Users Group, 2006).
54. Green, M. J., Girshkin, L., Kremerskothen, K., Watkeys, O. & Quidé, Y. A systematic review of studies reporting data-driven cognitive subtypes across the psychosis spectrum. *Neuropsychol. Rev.* **30**, 446–460 (2020).
55. Carruthers, S. P., Van Rheenen, T. E., Gurvich, C., Sumner, P. J. & Rossell, S. L. Characterising the structure of cognitive heterogeneity in schizophrenia spectrum disorders. A systematic review and narrative synthesis. *Neurosci. Biobehav. Rev.* **107**, 252–278 (2019).
56. Oomen, P. et al. The neurobiological characterization of distinct cognitive subtypes in early-phase schizophrenia-spectrum disorders. *Schizophr. Res.* **241**, 228–237 (2022).
57. Waltz, J. A., Frank, M. J., Robinson, B. M. & Gold, J. M. Selective reinforcement learning deficits in schizophrenia support predictions from computational models of striatal-cortical dysfunction. *Biol. Psychiat.* **62**, 756–764 (2007).
58. Morris, S. E., Heerey, E. A., Gold, J. M. & Holroyd, C. B. Learning-related changes in brain activity following errors and performance feedback in schizophrenia. *Schizophr. Res.* **99**, 274–285 (2008).
59. Farreny, A. et al. Study of positive and negative feedback sensitivity in psychosis using the wisconsin card sorting test. *Compr. Psychiatry* **68**, 119–128 (2016).
60. Ossola, P., Garrett, N., Biso, L., Bishara, A. & Marchesi, C. Anhedonia and sensitivity to punishment in schizophrenia, depression and opiate use disorder. *J. Affect. Disord.* **330**, 319–328 (2023).
61. Perry, W. & Braff, D. L. A multimethod approach to assessing perseverations in schizophrenia patients. *Schizophr. Res.* **33**, 69–77 (1998).
62. Lanser, M. G., Berger, H. J., Ellenbroek, B. A., Cools, A. R. & Zitman, F. G. Perseveration in schizophrenia: failure to generate a plan and relationship with the psychomotor poverty subsyndrome. *Psychiatry Res.* **112**, 13–26 (2002).
63. Waford, R. N. & Lewine, R. Is perseveration uniquely characteristic of schizophrenia?. *Schizophr. Res.* **118**, 128–133 (2010).
64. Park, S. & Gooding, D. C. Working memory impairment as an endophenotypic marker of a schizophrenia diathesis. *Schizophr. Res. Cogn.* **1**, 127–136 (2014).
65. Lee, J. & Park, S. Working memory impairments in schizophrenia: a meta-analysis. *J. Abnorm. Psychol.* **114**, 599 (2005).
66. Van Snellenberg, J. X. et al. Mechanisms of working memory impairment in schizophrenia. *Biol. Psychiat.* **80**, 617–626 (2016).
67. Fernyhough, C. & Borghi, A. M. Inner speech as language process and cognitive tool. *Trends Cogn. Sci.* (2023).
68. Baldo, J. V. et al. Is problem solving dependent on language?. *Brain Lang.* **92**, 240–250 (2005).
69. Ren, X., Wang, T. & Jarrold, C. Individual differences in frequency of inner speech: differential relations with cognitive and non-cognitive factors. *Front. Psychol.* **7**, 1675 (2016).
70. Li, C.-S.R. Do schizophrenia patients make more perseverative than non-perseverative errors on the wisconsin card sorting test? A meta-analytic study. *Psychiatry Res.* **129**, 179–190 (2004).
71. Tyburski, E. et al. Neuropsychological profile of specific executive dysfunctions in patients with deficit and non-deficit schizophrenia. *Front. Psychol.* **8**, 1459 (2017).
72. Gebreegziabhere, Y., Habatmu, K., Mihretu, A., Cella, M. & Alem, A. Cognitive impairment in people with schizophrenia: an umbrella review. *Eur. Arch. Psychiatry Clin. Neurosci.* **272**, 1139–1155 (2022).
73. Orellana, G. & Slachevsky, A. Executive functioning in schizophrenia. *Front. Psych.* **4**, 35 (2013).
74. Goldwin, M., Behar, E. & Sibrava, N. Concreteness of depressive rumination and trauma recall in individuals with elevated trait rumination and/or posttraumatic stress symptoms. *Cogn. Ther. Res.* **37**, 680–689 (2013).
75. Ehring, T., Frank, S. & Ehlers, A. The role of rumination and reduced concreteness in the maintenance of posttraumatic stress disorder and depression following trauma. *Cogn. Ther. Res.* **32**, 488–506 (2008).
76. Allen, D. N., Goldstein, G. & Warnick, E. A consideration of neuropsychologically normal schizophrenia. *J. Int. Neuropsychol. Soc.* **9**, 56–63 (2003).
77. Van Rheenen, T. E. et al. Characterizing cognitive heterogeneity on the schizophrenia-bipolar disorder spectrum. *Psychol. Med.* **47**, 1848–1864 (2017).
78. Hamm, J. A., Hasson-Ohayon, I., Kukla, M. & Lysaker, P. H. Individual psychotherapy for schizophrenia: trends and developments in the wake of the recovery movement. *Psychol. Res. Behav. Manage.* **2013**, 45–54 (2013).
79. Carbone, M. R. When not to use machine learning: a perspective on potential and limitations. *MRS Bull.* **47**, 968–974 (2022).
80. Liu, X., Lu, D., Zhang, A., Liu, Q. & Jiang, G. Data-driven machine learning in environmental pollution: gains and problems. *Environ. Sci. Technol.* **56**, 2124–2133 (2022).
81. Khaleghi, A., Mohammadi, M. R., Shahi, K. & Nasrabadi, A. M. Computational neuroscience approach to psychiatry: a review on theory-driven approaches. *Clin. Psychopharmacol. Neurosci.* **20**, 26 (2022).
82. Lim, T. V. & Ersche, K. D. Theory-driven computational models of drug addiction in humans: fruitful or futile?. *Addict. Neurosci.* **5**, 100066 (2023).
83. Gueguen, M. C., Schweitzer, E. M. & Konova, A. B. Computational theory-driven studies of reinforcement learning and decision-making in addiction: What have we learned?. *Curr. Opin. Behav. Sci.* **38**, 40–48 (2021).
84. Maia, T. V., Huys, Q. J. & Frank, M. J. Theory-based computational psychiatry. *Biol. Psychiat.* **82**, 382–384 (2017).

85. Huys, Q. J., Maia, T. V. & Frank, M. J. Computational psychiatry as a bridge from neuroscience to clinical applications. *Nat. Neurosci.* **19**, 404–413 (2016).
86. Berdia, S. & Metz, J. An artificial neural network stimulating performance of normal subjects and schizophrenics on the wisconsin card sorting test. *Artif. Intell. Med.* **13**, 123–138 (1998).
87. Monchi, O., Taylor, J. G. & Dagher, A. A neural model of working memory processes in normal subjects, parkinson's disease and schizophrenia for fmri design and predictions. *Neural Netw.* **13**, 953–973 (2000).
88. Amos, A. A computational model of information processing in the frontal cortex and basal ganglia. *J. Cogn. Neurosci.* **12**, 505–519 (2000).
89. Cella, M. et al. Identifying cognitive remediation change through computational modelling-effects on reinforcement learning in schizophrenia. *Schizophr. Bull.* **40**, 1422–1432 (2014).
90. Bishara, A. J. et al. Sequential learning models for the wisconsin card sort task: assessing processes in substance dependent individuals. *J. Math. Psychol.* **54**, 5–13 (2010).
91. Rathee, R., Luhrmann, T. M., Bhatia, T. & Deshpande, S. N. Cognitive insight and objective quality of life in people with schizophrenia and auditory hallucinations. *Psychiatry Res.* **259**, 223–228 (2018).
92. Barber, L., Reniers, R. & Upthegrove, R. A review of functional and structural neuroimaging studies to investigate the inner speech model of auditory verbal hallucinations in schizophrenia. *Transl. Psychiatry* **11**, 582 (2021).
93. Freeman, H. B. & Lee, J. Sex differences in cognition in schizophrenia: what we know and what we do not know. In *Cognitive Functioning in Schizophrenia: Leveraging the RDoC Framework* 463–474 (2022).
94. Zubicaray, G. D. & Ashton, R. Nelson's (1976) modified card sorting test: a review. *Clin. Neuropsychol.* **10**, 245–254 (1996).
95. Greve, K. W. The wcst-64: a standardized short-form of the wisconsin card sorting test. *Clin. Neuropsychol.* **15**, 228–234 (2001).
96. Schirner, M. et al. Brain simulation as a cloud service: the virtual brain on ebrains. *Neuroimage* **251**, 118973 (2022).
97. Appukuttan, S., Bologna, L. L., Schürmann, F., Migliore, M. & Davison, A. P. Ebrains live papers-interactive resource sheets for computational studies in neuroscience. *Neuroinformatics* **21**, 101–113 (2023).

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. This research has received funding from 'European Union - NextGenerationEU - PNRR', MUR code IR0000011, CUP B51E22000150006, project 'EBRAINS-Italy - European Brain ReseArch INfrastructureS Italy'.

Author contributions

GG: Idea, model design, model implementation, simulations, data analysis, result interpretation, writing. RC: model implementation, simulations, data analysis, result interpretation, writing AB: Idea, result interpretation, writing. AM: data analysis, result interpretation, writing. SC: simulations, data analysis, result interpretation, writing SR: data analysis, result interpretation, writing GB: Idea, result interpretation, writing, overall supervision. All authors reviewed the manuscript.

Competing interests

The authors have no financial, general, or institutional competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-89555-3.

Correspondence and requests for materials should be addressed to G.G.