frontiers | Research Topics

Computational models of brain in cognitive function and mental disorder

Edited by
Rubin Wang, Xu Lei, Jianzhong Su, Vito Di Maio and
Hans Albert Braun

Published in
Frontiers in Psychiatry

ISSN 1664-8714
ISBN 978-2-8325-4094-7
DOI 10.3389/978-2-8325-4094-7

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Computational models of brain in cognitive function and mental disorder

Topic editors

Rubin Wang — East China University of Science and Technology, China
Xu Lei — Southwest University, China
Jianzhong Su — University of Texas at Arlington, United States
Vito Di Maio — Institute of Applied Sciences and Intelligent Systems, Department of Physical Sciences and Technologies of Matter, National Research Council (CNR), Italy
Hans Albert Braun — University of Marburg, Germany

Citation

Wang, R., Lei, X., Su, J., Di Maio, V., Braun, H. A., eds. (2023). *Computational models of brain in cognitive function and mental disorder*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-4094-7

Table of contents

frontiers | Frontiers in Psychiatry

TYPE Editorial
PUBLISHED 27 November 2023
DOI 10.3389/fpsyt.2023.1230587

Check for updates

OPEN ACCESS

EDITED AND REVIEWED BY
Andreea Oliviana Diaconescu,
University of Toronto, Canada

*CORRESPONDENCE
Rubin Wang
✉ rbwang@163.com
Jianzhong Su
✉ su@uta.edu

RECEIVED 29 May 2023
ACCEPTED 13 November 2023
PUBLISHED 27 November 2023

CITATION
Wang R and Su J (2023) Editorial: Computational models of brain in cognitive function and mental disorder.
Front. Psychiatry 14:1230587.
doi: 10.3389/fpsyt.2023.1230587

Editorial: Computational models of brain in cognitive function and mental disorder

Rubin Wang[1]* and Jianzhong Su[2]*

[1]School of Mathematics, Institute for Cognitive Neurodynamics, East China University of Science and Technology, Shanghai, China, [2]Department of Mathematics, University of Texas at Arlington, Arlington, TX, United States

KEYWORDS

computational models, neuroscience, cognitive function, mental disorder, neurodynamics

Editorial on the Research Topic
Computational models of brain in cognitive function and mental disorder

The computational models in neuroscience are to utilize the modern computational tools to mimic brain behavior and to provide insight for the inner working of cognitive functions including their abnormal states in mental disorder. These cognitive models work at different levels of neuronal activities, from subcellular neuronal networks to full brain dynamics, and neuronal activates across different spatial and temporal scales is interactional and often ovelapping. It is vital for understanding and testing hypothesis in these models to understand how brain functions at normal cognitive processes as well as brain's diseased states in mental disorder patients.

The goal of this computational modeling Research Topic is to set a forum to enhance communication for quantitative explorations of the inner working of the brain cognitive dynamics, in both normal and pathological cognitive states. The authors study integrated network models at a single level or multiple levels of the brain, develop models of specific brain function or behavior, and simulate the models to mimic mechanisms of mental disorder from theoretical and computational methods.

This Research Topic covers a full range of Research Topics including the analysis of experimental data from cognitive neuroscience, cognitive disorder, mental disorder as well as theoretical models in neurodynamics using tools from mathematics and physics, computer science, etc.

This Research Topic included eight research papers and one review paper. In "*Category learning in a recurrent neural network with reinforcement learning*" by Zhang et al., the authors constructed a deep reinforcement neural learning model by combining a recurrent neural network (RNN) with reinforcement learning. They illustrated the category learning process and discussed how the machine learning process is represented in the neuron network. In "*Individual prediction of hemispheric similarity of functional connectivity during normal aging*" by Zhang, the author calculated the hemispheric functional connectivity (HSFC) through Pearson correlation of brain signals, then the author further evaluated the variability of individual recognition of HSFC during the aging process. In "*Gender differential item functioning analysis in measuring computational thinking disposition among secondary school students*" by Sovey et al., the research assessed gender's effects on students' ability in using computational thinking and the evaluate quantities include cognitive, affective, and conative dispositions. In "*Brain network changes in adult victims of violence*" by Shymanskaya et al., the authors compared brain network changes among two groups: self-identified victims of violence and the control group of individuals who did not identify themselves as victims. Four large-scale brain networks, the default mode network, the salience network, the fronto-parietal network, and the dorsal attention network were included in this study. In "*Turing instability mechanism of short-memory formation in multilayer FitzHugh-Nagumo network*" by Wang and Shen, a theoretical study is presented. The authors analyzed pattern properties of a network of neurons modeled by FitzHugh-Nagumo model, a relative simple model of oscillatory neuronal activity. Particularly, the authors studied neural activity patterns on a multilayer network where the coupled neuronal system is random network. In "*Dissociated deficits of anticipated and experienced regret in at-risk suicidal individuals*" by Ai et al., the authors studied subclinical youth with suicidal ideation and compared their quantitative behavior with a control group of youth who do not have suicidal ideation. They studied research subjects' responses in regret anticipation and experience during a value-based decision-making process. In "*Data-driven evolutionary game models for the spread of fairness and cooperation in heterogeneous networks*" by Li et al., the authors built evolutionary game models based on the experimental phenomena and data. Through the research, they showed a joint effect of social preference and network heterogeneity on promoting prosocial behaviors. In "*The distribution and heterogeneity of excitability in focal epileptic network potentially contribute to the seizure propagation*", by Fan et al., the authors used a connected network of neuronal units that have focal nodes prominently as their epilepsy model. Through computer simulation, they established a timescale difference that separated epileptic network model from non-seizure network. They concluded that factors affecting seizure occurrence are the connectivity patterns of focal network nodes and the network's ability to modulate the distribution of network excitability. Finally in a review paper "*Understanding mental health through computers: An introduction to computational psychiatry*", by Martínez and Santamaría-García, the authors performed a literature review for the field of computational psychiatry. They indicated the computational models have been established as a new tool in the study of mental disorders and problems. They suggested modeling integration by models of different neuronal levels will create computational phenotypes that are highly valued in clinical study and neuroscience research. They articulated that modeling study can be valuable in assisting physicians in precision psychiatry and the field of computational psychiatry has a strong potential to continue to grow as a new branch of computational neuroscience.

Author contributions

All authors contributed to manuscript revision, read, and approved the submitted version.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Psychiatry

TYPE Original Research
PUBLISHED 26 September 2022
DOI 10.3389/fpsyt.2022.1016807

OPEN ACCESS

EDITED BY
Vito Di Maio,
National Research Council (CNR), Italy

REVIEWED BY
Regina Júlia Deák-Meszlényi,
Hungarian Academy of Sciences
(MTA), Hungary
Bo Lu,
South China University of Technology,
China
Weijie Ye,
Guangdong University of Finance
and Economics, China

*CORRESPONDENCE
Yingteng Zhang
xiaoteng28@163.com

SPECIALTY SECTION
This article was submitted to
Computational Psychiatry,
a section of the journal
Frontiers in Psychiatry

RECEIVED 11 August 2022
ACCEPTED 31 August 2022
PUBLISHED 26 September 2022

CITATION
Zhang Y (2022) Individual prediction
of hemispheric similarity of functional
connectivity during normal aging.
Front. Psychiatry 13:1016807.
doi: 10.3389/fpsyt.2022.1016807

Individual prediction of hemispheric similarity of functional connectivity during normal aging

Yingteng Zhang*

Department of Mathematics, Taizhou University, Jiangsu Province, Taizhou, China

In the aging process of normal people, the functional activity pattern of brain is in constant change, and the change of brain runs through the whole life cycle, which plays a crucial role in the track of individual development. In recent years, some studies had been carried out on the brain functional activity pattern during individual aging process from different perspectives, which provided an opportunity for the problem we want to study. In this study, we used the resting-state functional magnetic resonance imaging (rs-fMRI) data from Cambridge Center for Aging and Neuroscience (Cam-CAN) database with large sample and long lifespan, and computed the functional connectivity (FC) values for each individual. Based on these values, the hemispheric similarity of functional connectivity (HSFC) obtained by Pearson correlation was used as the starting point of this study. We evaluated the ability of individual recognition of HSFC in the process of aging, as well as the variation trend with aging process. The results showed that HSFC could be used to identify individuals effectively, and it could reflect the change rule in the process of aging. In addition, we observed a series of results at the sub-module level and find that the recognition rate in the sub-module was different from each other, as well as the trend with age. Finally, as a validation, we repeated the main results by human brainnetome atlas (BNA) template and without global signal regression, found that had a good robustness. This also provides a new clue to hemispherical change patterns during normal aging.

KEYWORDS

hemispheric similarity of functional connectivity, functional MRI, normal aging, individual recognition, global signal

Introduction

In recent years, a large number of studies (1–3) have used the combination of pattern recognition and brain image data to distinguish healthy elderly people from Alzheimer's disease (AD) patients, and achieve good results. In addition to using structural MRI (sMRI) data to explore cortical atrophy and white matter fiber tracts abnormalities in specific areas of AD, several studies (4, 5) have also used fMRI data to explore differences in brain functional activity between healthy elderly people and AD patients. The above researches reflect the distribution pattern of brain structure and function in people with abnormal aging (i.e., suffering from common nervous system diseases such as AD). However, in the life cycle of normal people, from youth, middle age to old age, the pattern of brain functional activity is constantly changing. There is a lack of relevant research on the change rule with the aging process, which has always been the focus of attention in the field of cognitive neuroscience.

To investigate the difference pattern of individual brain functional activity during normal aging, some scholars (6–8) study a series of metrics derived from fMRI, such as regional homogeneity (ReHo), amplitude of low frequency fluctuation (ALFF) and functional connectivity (FC). Most of these indicators are studied at the whole brain level, some specific regions of interest (ROI) or homologous brain regions, and the correlation of activity patterns between the left and right hemispheres is not clear. In recent years, pattern recognition has been applied more and more widely in neuroimaging and numerous individual recognition methods are constantly innovating. For example, Finn et al. (9) used the rs-fMRI and task fMRI (tfMRI) data of a large sample from human Connectome Project (HCP) in 2015. Their research demonstrated that functional connectivity, as a kind of "fingerprinting," could effectively identify individuals from large samples, and that the sub-network with the most significant difference among individuals could well predict individual differences in fluid intelligence. In addition, Kaufmann et al. (10) used this fingerprinting method in 2017 to show that delayed brain network development during adolescence was associated with decreased mental health. However, the effectiveness of this "fingerprinting" approach in identifying individuals during normal aging remains unclear.

Brain changes occur throughout the life cycle and play a critical role in individual developmental trajectories for cognition, social functioning, adaptability, personality and mental health. Due to the great potential of neuroplasticity and the continuous development of environmental sensitivity, some scholars hypothesize that functional connectivity shapes individual differences in individual maturation and aging mechanisms. In recent years, several studies (11–13) have made use of Cam-CAN database to study the brain functional activity pattern of individual aging process from different perspectives, which provide an opportunity for our research.

Here, we proposed the metric of the left and right hemispheric similarity of functional connectivity (HSFC) to explore whether the hemispheric similarity had the characteristics of individual differences in groups of different ages and how it changed during aging. In particular, we used the Cam-CAN dataset for a population aged 18–88 years and constructed hemispheric functional connectivity networks for rs-fMRI data of each individual. Then, the HSFC computed by Pearson correlation was used as the starting point of this study to evaluate the individual identification ability of HSFC in the aging process and its correlation with age. In addition, we observed a series of results of HSFC at the sub-module level. Finally, as a validation, we repeated the main results through another functional template and no global signal regression (NGSR).

Materials and methods

Subjects

The Cam-CAN Stage 2 dataset[1] (14) included 646 subjects with T1 and rs-fMRI data (age range: 18∼88 years, 314 males) was used. All the subjects were native English speakers, had normal or corrected vision and hearing, scored 25 or above on the mini-mental state examination (MMSE), and had no neurological disorders. It was worth noting that 4 subjects are excluded from this dataset due to incomplete data collection. Thus, a total of 642 subjects entered the preprocessing step. Ethical approval was approved by the University of Cambridge's Research Ethics Committee. All subjects gave written informed consent.

All scans were performed using the standard 3T Tim Trio (Siemens) with 32 channel coils. The rs-fMRI scans were obtained using EPI sequences: whole brain coverage; 261 volumes, each volume contains 32 axial slices; layer thickness 3.7 mm with an 20% inter-slice gap; TR = 1,970 ms; TE = 30 ms; FOV = 192 × 192 mm^2; flip angle = 78°; voxel size = 3 × 3 × 4.44 mm^3. High resolution T1-weighted structure images were obtained using MPRAGE sequence, and the parameters were as follows: TR = 2,250 ms; TE = 2.99 ms; TI = 900 ms; FOV = 256 × 240 × 192 mm^3; flip angle = 9°; voxel size = 1 mm; isotropy; generalized automatic calibration partial parallel acquisition (GRAPPA) acceleration factor = 2.

Data processing

Firstly, using the FUGUE tool of the FSL package to accomplish the fieldmap correction.[2] According to the phase difference image and short TE amplitude images to get rad images and then used the rad images of EPI image correction. Then DPABI toolbox was used to preprocess the resultant rs-fMRI images (15), including the following steps: ① removed the

[1] http://www.cam-can.org/

[2] https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FUGUE

first 10 time points; ② time layer correction; ③ head movement correction; ④ the diffeomorphic anatomical registrations through exponentiated lie algebra (DARTEL) (16) segmentation method dealt with sMRI scans and used it to normalize rs-fMRI scans. ⑤ standardization; ⑥ regression covariables (including Friston's 24 head movement parameters (17), global signal, average signal of white matter and cerebrospinal fluid); ⑦ bandpass filtering (0.01–0.1 Hz). It was worth noting that there was some controversy over whether the global signal should be regressed during rs-fMRI data preprocessing (18–20). In 2009, there existed opposite recommendations about whether GSR should be used in the processing of rs-fMRI data (18, 19). Murphy et al. was the first to show that GSR mathematically mandates the presence of anti-correlations (19). Because anti-correlations following GSR could be an artifact of the processing technique, Murphy et al. concluded that GSR should not be used. However, Fox et al. found that several characteristics of anti-correlated networks could not be attributed to GSR. Because GSR enhanced the detection of system-specific correlations and improved the correspondence between resting-state correlations and anatomy, they concluded that GSR can be beneficial (18). Therefore, we also calculated NGSR in the step of regression covariate to explore the influence of global signal on the results. In the preprocess, 14 subjects with a head movement of more than 3 mm and 3° and 1 subject with segmentation failure were removed. A total of 627 subjects were included in the analysis. There were 166 subjects in the Young group (18∼39 years old), 197 subjects in the Middle group (40∼59 years old), and 264 subjects in the Old group (60∼88 years old). The information of subjects was shown in **Table 1**. There was no significant difference in gender ($P = 0.871$) and significant difference in age ($P < 0.0001$). The statistical analysis of basic information was obtained through SPSS22.0.

Constructing functional network

The construction process of functional network was shown in **Figure 1A**. For Cam-CAN data, we used the atlas of intrinsic connectivity of homotopic areas (AICHA) (21) to extract the average time series of each ROI. The atlas divided the brain into 384 regions (192 regions in each hemisphere), containing 344 cortical regions and 40 subcortical regions. It had been used in some studies to divide the brain for FC and brain network analysis (22–24). For each subject, we obtained the mean time series of 384 regions through the time series of all voxels in each ROI. The FC between two brain regions was obtained by calculating Pearson correlation coefficients of average time series. Finally, each subject obtained a 384 × 384 symmetric FC matrix. Each intra-hemisphere network was a 192 × 192 symmetric FC matrix and had been used Fisher-z transform to make the statistical normalization.

In order to explore the contribution of different ROI to individual recognition, we further subdivided the hemispheric functional network into five sub-modules (i.e., heteromodal, paralimbic, primary, unimodal and subcortical) based on functional hierarchy (25). This functional hierarchy was based on studies of anatomy, electrophysiology, behavior, injury, and functional imaging in non-human primates and humans. The heteromodal and unimodal areas were most closely involved in perceptual elaboration and motor planning. The paralimbic areas played a critical role in channeling emotion and motivation to behaviorally relevant intrapsychic and extrapersonal targets. The primary included primary sensory cortex, primary motor cortex, primary visual cortex, primary auditory cortex, primary somatosensory cortex and primary gustatory cortex and these cortices mainly responsible for the control of motor, visual processing, auditory processing and other functions. The subcortical included insula, amygdala, putamen and thalamus. Among them, the thalamus relays communication among subcortical and cortical regions and played a central role in the integration of sensory information. The cortical distribution of the five sub-modules was shown in **Figure 1C**. Many studies had used these sub-modules (26–28).

Individual identification steps for hemispherical functional networks

The individual identification method used in this paper was a reference to the work of Finn et al. (9). Finn et al. used the rs-fMRI and tfMRI data from HCP database and this research demonstrated that functional connectivity, as a kind of "fingerprinting," could effectively identify individuals from large samples. The difference between the individual recognition of Finn et al. and ours was that Finn et al. computed the Pearson correlation between the functional connectivity of the whole brain of an individual and the functional connectivity of the whole brain of another scan, while we computed

TABLE 1 Subject demographics.

	Young	Middle	Old	*P*-value
Sample size	166	197	264	
Gender (male/female)	79/87	95/102	132/132	0.871
Age (years)	30.56 ± 5.68	49.21 ± 5.67	72.71 ± 7.53	<0.0001

The ages are shown as mean ± standard deviation (SD). Columns on the right display *P*-value by *F*-test for age and the gender computes *P*-value by chi-square test.

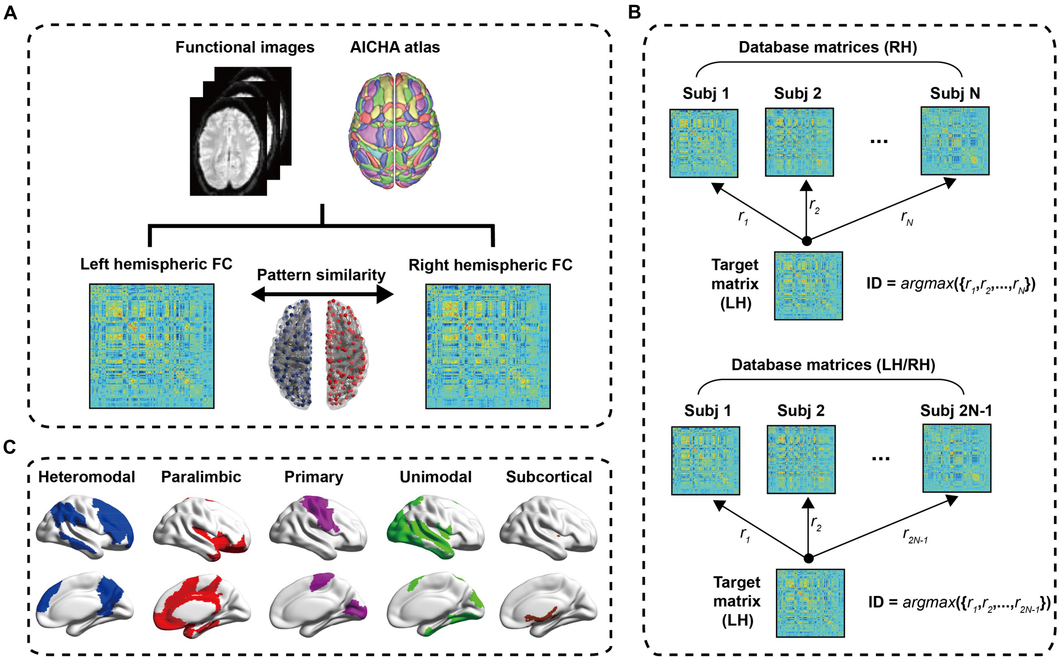


FIGURE 1
The process flow chart. **(A)** Data preprocessing. After a series of preprocessing steps, a 384 × 384 symmetric resting state FC matrix is obtained, and the left hemisphere and right hemisphere is 192 × 192 symmetric connection matrices, respectively. The Pearson correlation of left hemisphere and right hemisphere of intra-subject is defined as the hemispheric similarity of functional connectivity (HSFC). **(B)** Schematic diagram of individual identification; **(C)** the cortical distribution of the five sub-modules. LH, Left hemisphere; RH, Right hemisphere.

the Pearson correlation between the functional connectivity of the left and right intra-hemispheres of an individual to complete the recognition process. Figure 1B showed the process of the LH recognizing the RH in individual. First, created database matrices containing right hemisphere FC matrices for all subjects. $D = [X_i, i = 1, 2, \cdots, N]$, X_i was a 192 × 192 FC matrix, Subscript i refered to the subject, N represented the total number of subjects. In the identification step, the similarities between the target matrix and all the right hemisphere FC matrices in the dataset were calculated. These similarities were defined as Pearson correlation between the target matrix and each FC matrix in the dataset. When the target matrix (LH) and a matrix (RH) in the dataset obtained the maximum Pearson correlation value and their ID was the same $[ID = argmax\ (\{r_1,\ r_2, \cdots,\ r_N\})]$, it meant correct identification. The upper part of the dataset matrices in Figure 1B were the FC matrices of RH, that was, the contralateral hemisphere was used as a test set to identify individual. And the lower part of Figure 1B also contained all FC matrices of LH except the target matrix, namely using ipsilateral and contralateral hemisphere as a test set to identify individual. Similarly, the steps of the RH to recognize the LH were consistent with the above process. In order to evaluate the validity and robustness of this identification method in statistics, a non-parametric permutation test was performed. In each recognition process, we randomly shuffled the subjects' hemispheres in the dataset, and then used each target matrix to identify them in turn, and compared the difference between the obtained recognition rate and the initial recognition rate. This process

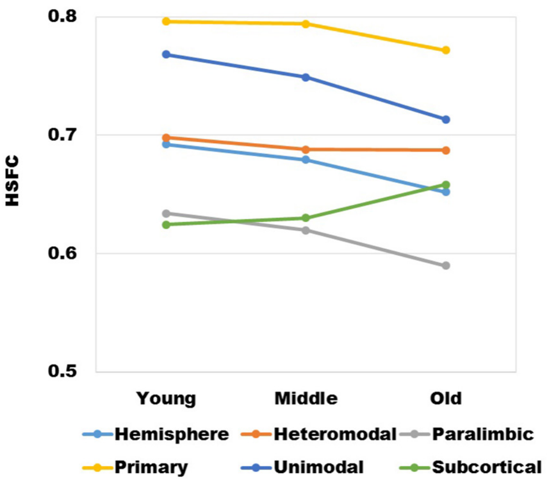


FIGURE 2
The variation trend of HSFC of hemisphere and each sub-module (heteromodal, parallel, primary, unimodal, subcortical) in the process of aging.

was performed 1,000 times. In order to explore the contribution of sub-modules to individual recognition, we carried out individual recognition for each of the five sub-modules, and the recognition steps were basically the same as the hemispheric recognition process. In the following content, we also defined the hemispheric similarity of each sub-module as HSFC.

Age-related changes in hemispheric similarity

A large number of studies (29–32) had shown that aging could affect the FC between brain regions, not only the connectivity within functional subnetworks, but also the connectivity between different functional subnetworks. Aging caused the brain networks of older people to become less modular, as well as reduced local efficiency. In order to investigate the variation trend of HSFC during aging, we calculated Pearson correlation between subjects' age and HSFC. At the same time, the above operations were also performed on five sub-modules.

Validation analysis

In this study, individual identification and the relationship between HSFC and age were conducted based on AICHA template. In order to explore the stability of the calculation results for atlas, we used the human brainnetome atlas (BNA)[3] for validation analysis (33). The BNA was based on a connective architecture that allowed brain anatomy to be correlated with psychological and cognitive functions and therefore was suitable for functional brain network analysis. The atlas divided the brain into 246 regions (123 for each hemisphere), comprising 210 cortical regions and 36 subcortical regions. It had been used in some studies to divide the brain for FC and brain network analysis (34–37). For each subject, referring to AICHA's FC matrix construction process, finally we got a 246 × 246 symmetric FC matrix. Each intra-hemisphere was 123 × 123 symmetric FC matrix. The above AICHA's results were repeated using the FC matrix obtained by the BNA. At the same time, we compared the robustness of HSFC between different templates.

[3] http://atlas.brainnetome.org/

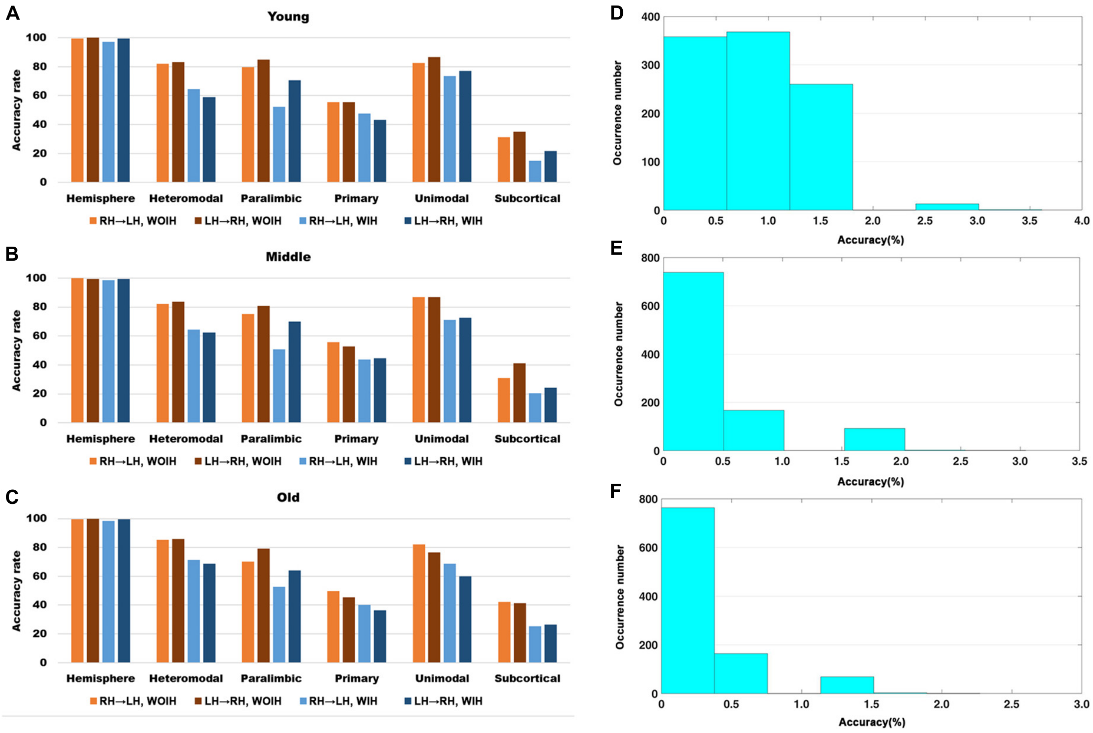


FIGURE 3
The recognition rate results of hemisphere and each sub-module in different age groups. **(A–C)** Represents the recognition rate of young, middle and old, respectively. From left to right in each sub graph, the recognition rate of hemisphere and five sub-modules are in turn. Notably, Orange indicates that the RH recognizes the LH without Ipsilateral Hemisphere (RH→LH, WOIH). Brown indicates that the LH recognizes the RH without ipsilateral hemisphere (LH→RH, WOIH). Light blue indicates that the RH recognizes the LH with ipsilateral hemisphere (RH→LH, WIH). Dark blue indicates that the LH recognizes the RH with ipsilateral hemisphere (LH→RH, WIH). **(D–F)** Correspond to non-parametric permutation test of **(A–C)**, respectively.

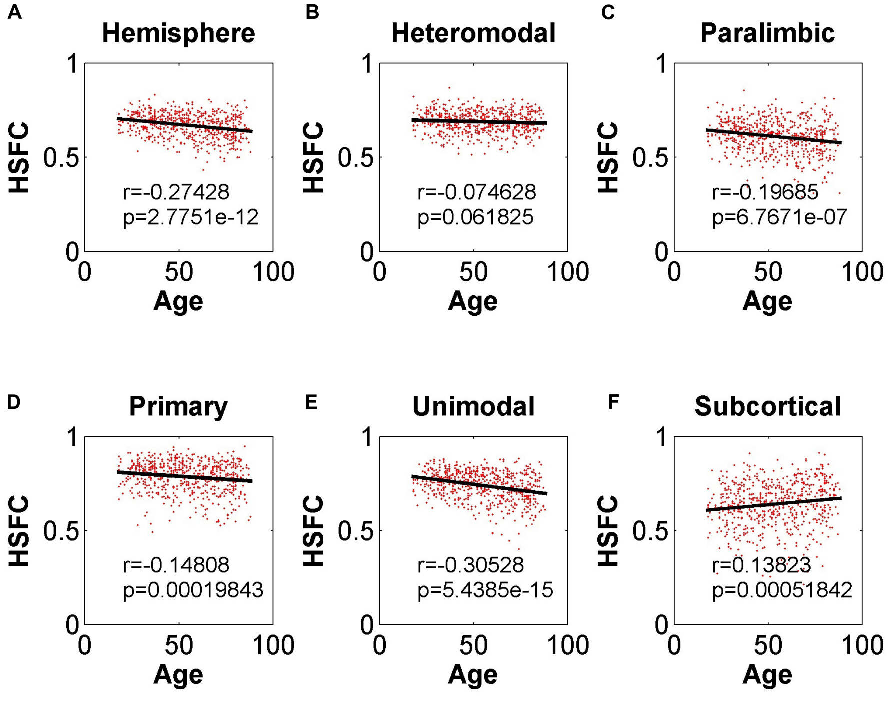


FIGURE 4

Pearson correlation between age and HSFC of hemisphere and five sub-modules. **(A–F)** Represent the hemisphere, heteromodal, paralimbic, primary, unimodal, subcortical, respectively.

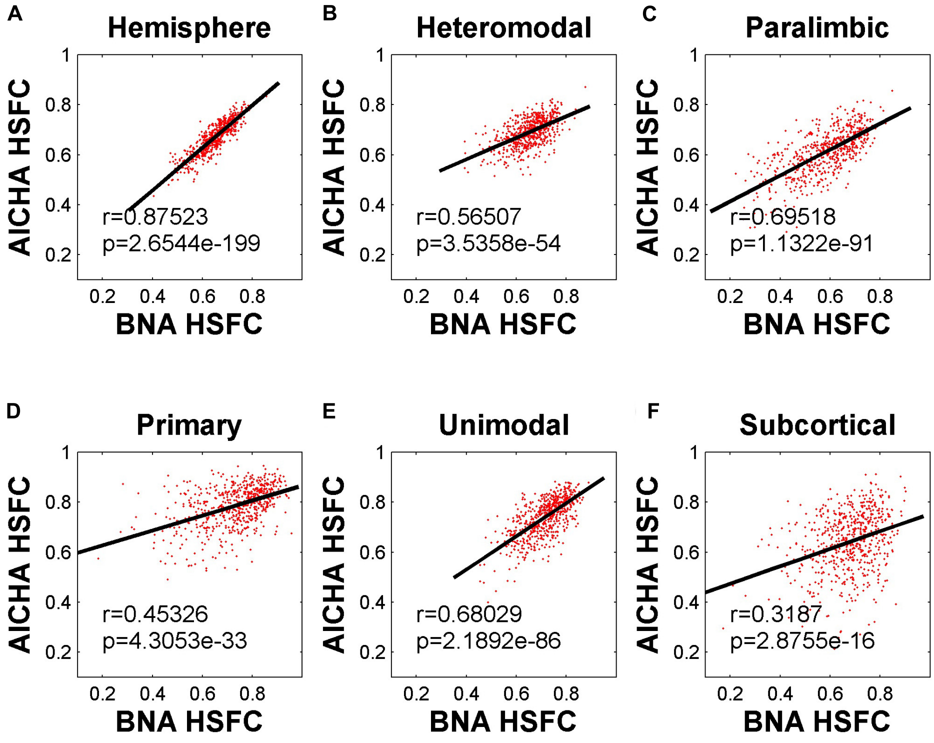


FIGURE 5

Pearson correlation of HSFC of hemisphere and five sub-modules between BNA and AICHA. **(A–F)** Represent the hemisphere, heteromodal, paralimbic, primary, unimodal, subcortical, respectively.

In addition, the above analysis was repeated with NGSR to explore the effect of global signal on the results.

Results

Hemispheric similarity of functional connectivity of hemispheres and sub-modules in different age group

It could be seen from Figure 2 that with the increase of age, except for subcortical, HSFC in other sub-modules and hemispheric level showed a decreasing trend, and heteromodal had the smallest decline. For different age group, the HSFC of primary always maintained the maximum value, followed by unimodal and heteromodal. In youth and middle age, the HSFC of unimodal was higher than that of subcortical, while in old age, the HSFC of subcortical was slightly higher than that of unimodal. In addition, the HSFC of paralimbic was slightly larger than that of subcortical in youth. With the aging process, the HSFC of paralimbic continues to decline, and the gap between paralimbic and subcortical was growing.

Individual recognition of hemispheric similarity of functional connectivity

We first observed the individual recognition results without ipsilateral hemisphere from Figure 3. It could be found that the individual recognition results of different age groups were roughly the same. Among them, the recognition ability of hemispheric level was the best and that of subcortical was the lowest. Heteromodal, paralimbic and unimodal had similar recognition abilities, which were slightly higher than primary. After adding the ipsilateral hemisphere for recognition, the recognition ability of each sub-module decreased to varying degrees, while the hemisphere level recognition had little effect. In addition, for the difference of LH to recognize RH or RH to recognize LH, there was little difference at the hemispheric level, but there were partial differences in different sub-modules. Given that the identification trials were not independent from one another, we performed non-parametric permutation testing to assess the statistical significance of these results. Across 1,000 iterations, the highest success rates achieved were 6/166 (Young group), 6/197 (Midlle group),6/264 (Old group), neither of which exceeded 4%. Thus the *P*-value associated with obtaining at least correct identifications (the minimum rate we achieved) was 0.

Age-related changes in hemispheric similarity

As shown in Figure 4, except for the positive correlation between HSFC and age in subcortical, the hemispheric and other sub-modules reflected the negative correlation trend. In addition, except that the correlation between HSFC and age was not significant ($r = -0.075$, $p = 0.06 > 0.05$) in heteromodal, HSFC of other sub-modules and hemispheric showed a significant correlation with age to varying degrees.

Validation analysis of template and processing method

The HSFC used in the previous main work was based on AICHA template. In order to understand whether the HSFC was specific to AICHA template, we recalculated HSFC using BNA template, and then computed Pearson correlation on the HSFC obtained from the two templates. As shown in Figure 5,

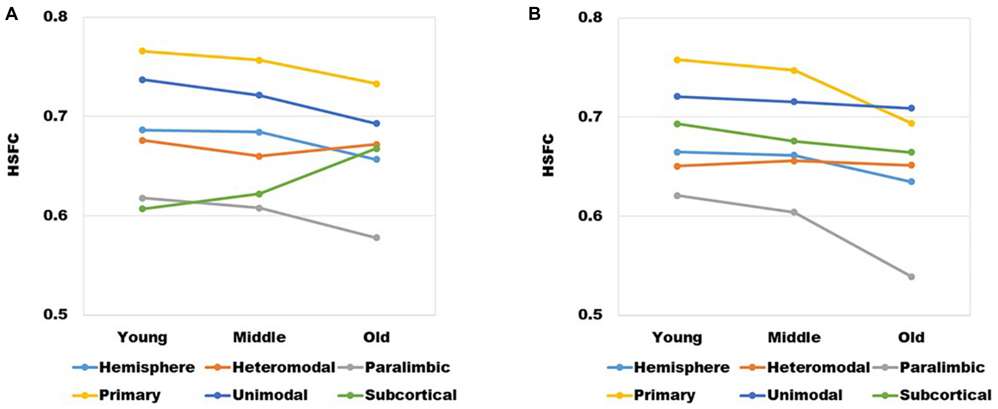


FIGURE 6
(A) The changes of HSFC with GSR obtained by BNA and **(B)** the changes of HSFC with NGSR obtained by AICHA in different age groups and modules.

the HSFC between templates showed a very significant positive correlation (r = 0.88, p = 2.65 × 10^{-199}) at the hemispheric level. The degree of correlation varied for different sub-modules. The correlation of subcortical was the lowest and the data points fitting were not strong.

The mean value of HSFC obtained by BNA was shown in Figure 6A. It seemed some differences when compared with the HSFC obtained by AICHA (Figure 2). Except for hemisphere and heteromodal, the mean value and change trend of other modules were basically similar to the HSFC obtained by AICHA. The HSFC with NGSR obtained by AICHA was shown in Figure 6B and the HSFC distribution was different from that obtained by GSR (Figure 2). Except for heteromodal, the mean value of HSFC in other modules decreased with aging. In

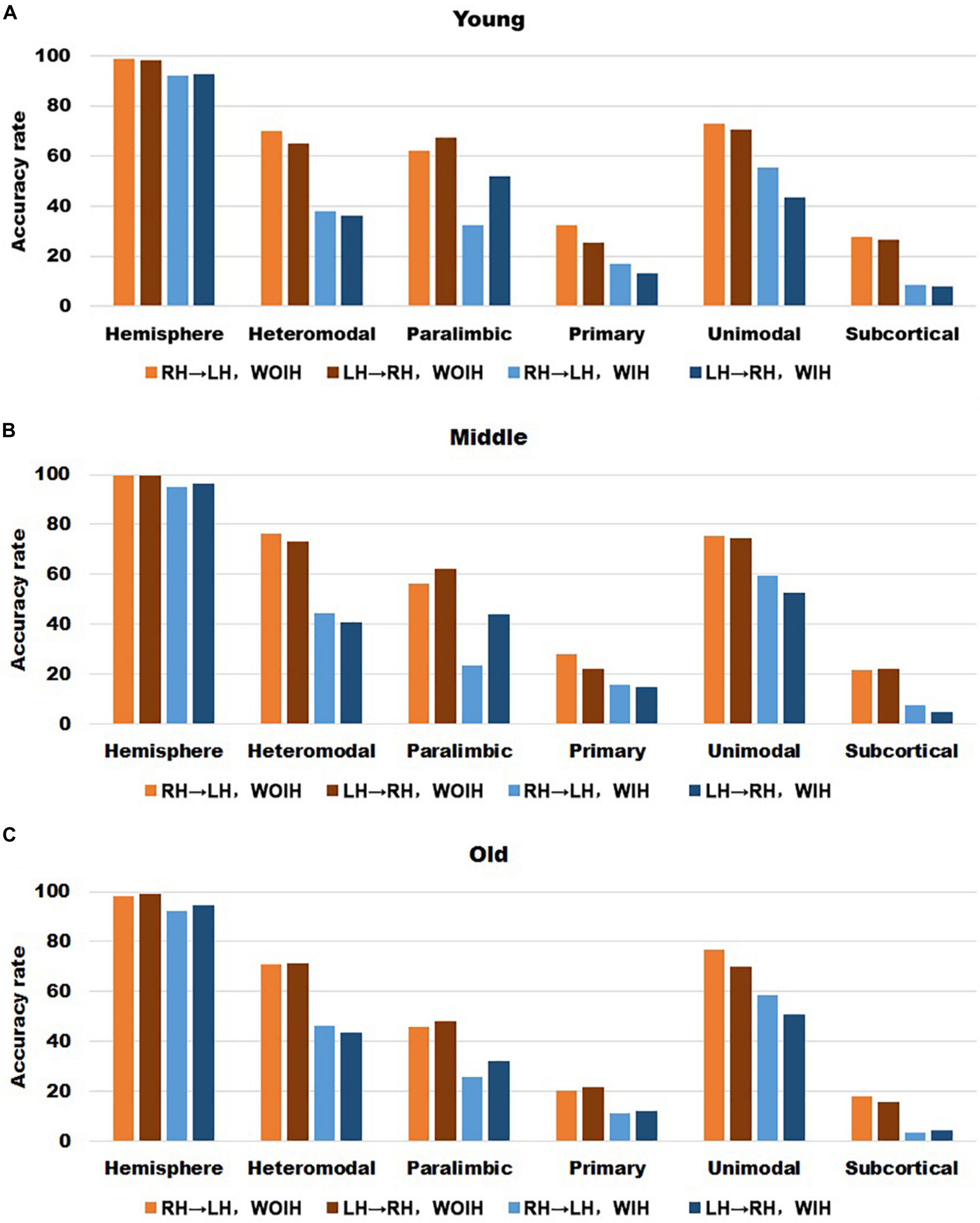


FIGURE 7
The recognition rate results of hemisphere and each sub-module in different age groups for BNA template with GSR. **(A–C)** Represents the recognition rate of young, middle and old, respectively. From left to right in each sub graph, the recognition rate of hemisphere and five sub-modules are in turn. Notably, orange indicates that the RH recognizes the LH Without Ipsilateral Hemisphere (RH→LH, WOIH). Brown indicates that the LH recognizes the RH Without Ipsilateral Hemisphere (LH→RH, WOIH). Light blue indicates that the RH recognizes the LH With Ipsilateral Hemisphere (RH→LH, WIH). Dark blue indicates that the LH recognizes the RH With Ipsilateral Hemisphere (LH→RH, WIH).

addition, the value of subcortical was higher than that of GSR and was greater than that of heteromodal in different age groups, while the HSFC of unimodal was higher than that of primary in old age. However, the contrast of HSFC between hemisphere and heteromodal in youth and middle age was opposite to that with GSR.

In the validation part of recognition rate, first of all, we observed the recognition rate results obtained from the BNA template (Figure 7). The individual recognition rates of paralimbic, primary and subcortical in the elderly were lower than those in the youth and middle age. In the comparison of the recognition rate of different templates, it was found that the recognition rate of AICHA template was better than that of BNA template, especially at the sub-module level. Next, after comparing the recognition rate results of GSR (Figure 3) and NGSR (Figure 8), we could find that the recognition rate of each module was almost the same.

In the validation part of the correlation between HSFC and age, similarly, we used the BNA template for validation (Figure 9) and found that the distribution patterns between the two templates were similar. The Pearson correlation between the HSFC and age for heteromodal and unimodal had no significant difference (heteromodal: $r = -0.002$, $p = 0.96$; unimodal: $r = -0.043$, $p = 0.28$). The correlation coefficient obtained by primary and paralimbic was larger than that of the corresponding sub-module in AICHA template. In addition, the correlation value between HSFC and age of primary and subcortical with NGSR (Figure 10) was higher than that with GSR, and the other modules were the opposite.

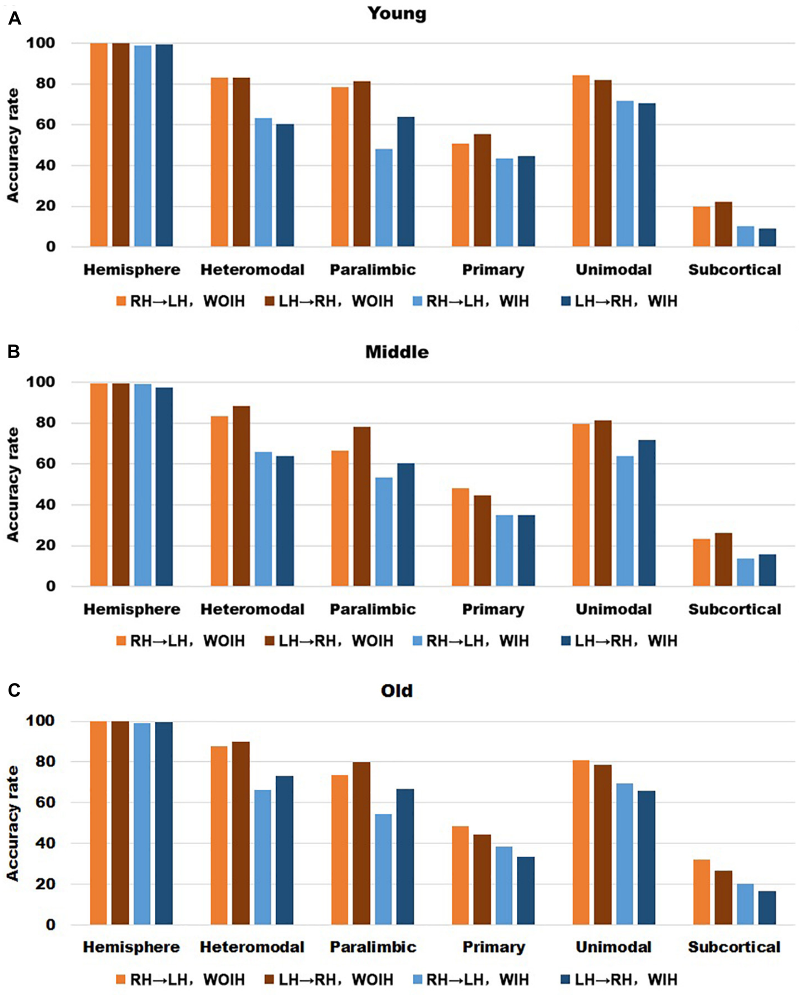


FIGURE 8
The recognition rate results of hemisphere and each sub-module in different age groups for AICHA template with NGSR. **(A–C)** Represents the recognition rate of young, middle and old, respectively. From left to right in each sub graph, the recognition rate of hemisphere and five sub-modules are in turn. Notably, orange indicates that the RH recognizes the LH Without Ipsilateral Hemisphere (RH→LH, WOIH). Brown indicates that the LH recognizes the RH Without Ipsilateral Hemisphere (LH→RH, WOIH). Light blue indicates that the RH recognizes the LH With Ipsilateral Hemisphere (RH→LH, WIH). Dark blue indicates that the LH recognizes the RH With Ipsilateral Hemisphere (LH→RH, WIH).

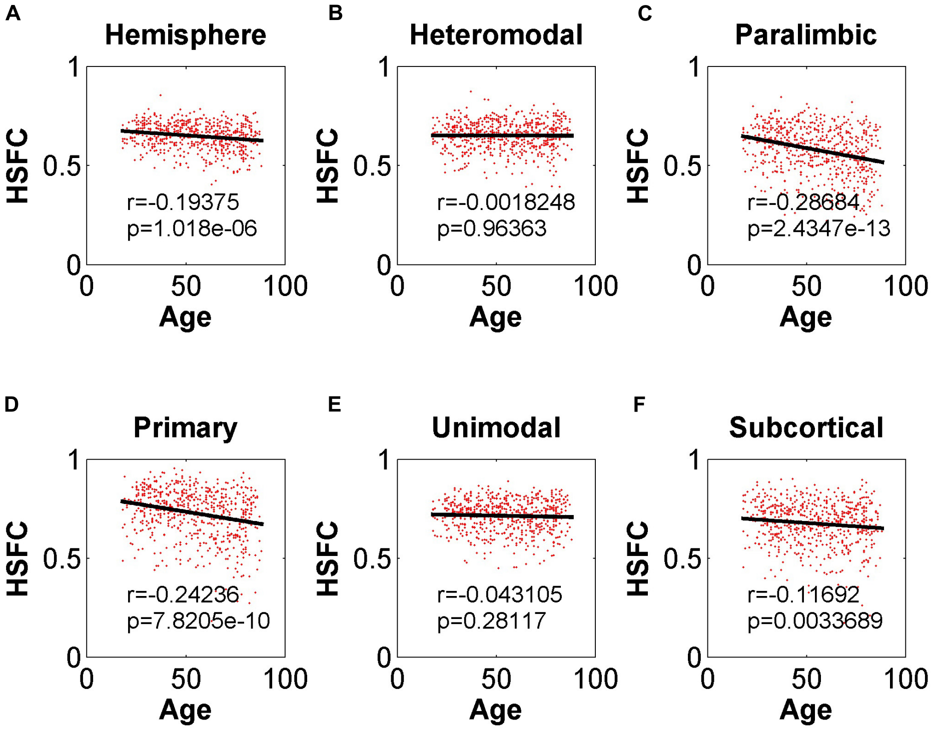


FIGURE 9

Pearson correlation between age and HSFC of hemisphere and five sub-modules for BNA template with GSR. **(A–F)** Represent the hemisphere, heteromodal, paralimbic, primary, unimodal, subcortical, respectively.

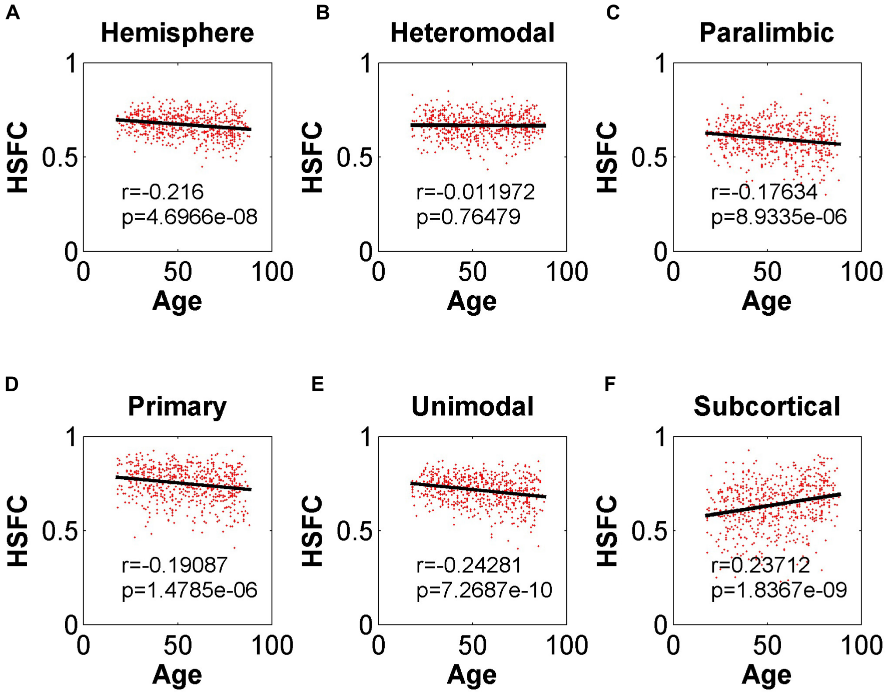


FIGURE 10

Pearson correlation between age and HSFC of hemisphere and five sub-modules for AICHA template with NGSR. **(A–F)** Represent the hemisphere, heteromodal, paralimbic, primary, unimodal, subcortical, respectively.

Discussion

In this study, the Cam-CAN datasets with a large age span were used and the index named "hemispheric similarity of functional connectivity (HSFC)" was proposed. This index could effectively identify individuals and reflect the change trend in the aging process. In addition, the results obtained in different sub-modules were also different. The results were robust to different templates and whether the global signal was regressed or not. This proves that HSFC has unique advantages in aging research, and also shows that HSFC has the characteristics of individual differences.

The specificity of the cerebral hemisphere is a sign of successful neural development (38). Previous studies (39–41) extracted a series of indicators as features through the specificity of hemispheric function or the asymmetry of hemispheric structure and function, and obtained a high accuracy in the diagnosis of diseases. On the contrary, some studies (42, 43) found the conclusion of hemispheric asymmetry through the processing and statistical analysis of imaging data. The above studies indirectly revealed the importance of the cerebral hemisphere, suggesting the starting point of this study.

In the past, the application of pattern recognition in imaging research was generally by extracting the features of different levels of the brain and building a classifier for the prediction of category variables or building a regressor for the continuous variables of behavior scores. Next, using a new test set to get the results on the classifier or regressor. Different from the common pattern recognition methods, this study based on the "fingerprint" method proposed by Finn et al. (9) had achieved a very high accuracy at the hemispheric level, which showed that each individual is unique. For sub-modules, since the primary module mainly involves primary cortical areas such as the central gyrus (25), the FC similarity of homologous brain areas between hemispheres is also very high and the degree of lateralization is small. Therefore, its HSFC value was the highest among all modules (**Figure 2**). Meanwhile the functional patterns of the primary module in the LH and RH are very similar so that the individual differences at the group level are not high, which led to a low recognition rate (**Figure 3**). Subcortical module mainly involves subcutaneous nuclei (25). The segmentation effect of subcutaneous nuclei in image data preprocessing is poor, which also indirectly affects the calculation of HSFC, resulting in its generally low value. Therefore, the individual recognition ability was not strong. For the difference of recognition rate in sub-modules, we hypothesized that this might be due to differences in functional connectivity similarities between homologous brain regions of different modules, leading to differences in the degree of lateralization, and thus affecting HSFC. We believe that HSFC can better reflect the degree of lateralization in different brain regions. The higher the value of HSFC, the higher the similarity of functional connectivity of homologous brain regions in this region, and the smaller the degree of lateralization. The smaller the value of HSFC is, the lower the similarity of functional connections of homologous brain regions in this region, and the greater the degree of lateralization. This can help us further explore differences in the degree of lateralization in different regions of the brain. In different age groups, the results of recognition rate were basically the same, which also showed that the individual differences of HSFC were stable in the aging process and had good robustness.

In the previous study (13), it was found that the changes of vascular components, head movements and the location of functional areas would affect the relevant patterns of FC and aging process, so a series of analysis and processing methods were proposed. Another study (44) showed that the shrinkage rate of various regions of the cerebral cortex during aging was not the same. In this study, based on the relationship between the HSFC and age, we found that HSFC decreased with the aging process on the whole. The results showed that the aging process led to the pattern disorder of many functional subnetworks, which disrupted the symmetry of hemispheric functional networks to some extent and further provided valuable clues for the future study of the development pattern of hemispheric functional networks in the aging process.

Through the study of the HSFC between different templates, it showed that HSFC is not only specific to a fixed template, but also could be extended to more functional templates. When using BNA template or NGSR, the results obtained were basically consistent with our main results (i.e., GSR with AICHA template).

In the outlook of the follow-up work, first of all, the FC network of this study was calculated by Pearson correlation. Some studies (45, 46) proposed the processing strategy of "distance correlation" and its research results were better than Pearson correlation, which was worthy of our reference in the future. Second, although this study used a wide range of aging data, it was limited to rs-fMRI research. In the future, structural MRI and task fMRI can be added for a more comprehensive analysis or we can consider applying the HSFC-based method to HCP datasets with different scans, so as to verify the recognition stability of HSFC at different time points in the same individual. Third, this study was aimed at a series of conclusions obtained in the process of normal aging, and its application prospect in Alzheimer's disease and other nervous system and mental diseases is not clear. Fourth, the continuous optimization of preprocessing strategy and the realization of large sample data are still big problems that have been committed to research in the field of pattern recognition, which still need to be solved.

In this study, the HSFC was proposed for the first time and it could effectively identify individuals and reflected the changes in the aging process. In particular, we found that there are differences in the recognition rate among sub-modules and there were also differences in the trend with age. Finally, as a validation, we repeated the main results through another functional template and NGSR, which had good robustness. This also provides new clues for the pattern of changes between hemispheres in the normal aging process.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: http://www.cam-can.org/.

Ethics statement

The studies involving human participants were reviewed and approved by the University of Cambridge's Research Ethics Committee. The patients/participants provided their written informed consent to participate in this study.

Author contributions

The author confirms being the sole contributor of this work and has approved it for publication.

Funding

This research was supported by the Natural Science Foundation of the Jiangsu Higher Education Institutions of China (No. 22KJB310019), the Scientific Research Foundation of High-level Personnel of Taizhou University (No. TZXY2021QDJJ001), the "2022 Taizhou Tuoju Project" for Young Science and Technology Talents, the National Natural Science Foundation of China (No. 12202208), and the Project of Excellent Science and Technology Innovation Team of Taizhou University.

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Battineni G, Hossain M, Chintalapudi N, Chintalapudi N, Traini E, Dhulipalla V, et al. Improved Alzheimer's disease detection by MRI using multimodal machine learning algorithms. *Diagnostics.* (2021) 11:2103. doi: 10.3390/diagnostics11112103

2. Pang Z, Wang X, Wang X, Qi J, Zhao Z, Gao Y, et al. A multi-modal data platform for diagnosis and prediction of Alzheimer's disease using machine learning methods. *Mobile Netw Appl.* (2021) 26:2341–52. doi: 10.1007/s11036-021-01834-1

3. Goyal P, Rani R, Singh K. State-of-the-art machine learning techniques for diagnosis of Alzheimer's disease from MR-images: A systematic review. *Arch Comput Methods Eng.* (2022) 29:2737–80. doi: 10.1007/s11831-021-09674-8

4. Shi Y, Zeng W, Deng J, Nie W, Zhang Y. The identification of Alzheimer's disease using functional connectivity between activity voxels in resting-state fMRI data. *IEEE J Transl Eng Health Med.* (2020) 8:1–11. doi: 10.1109/JTEHM.2020.2985022

5. Yang F, Jiang X, Yue F, Wang L, Boecker H, Han Y, et al. Exploring dynamic functional connectivity alterations in the preclinical stage of Alzheimer's disease: An exploratory study from SILCODE. *J Neural Eng.* (2022) 19:016036. doi: 10.1088/1741-2552/ac542d

6. Bernier M, Croteau E, Castellano CA, Cunnane SC, Whittingstall K. Spatial distribution of resting-state BOLD regional homogeneity as a predictor of brain glucose uptake: A study in healthy aging. *Neuroimage.* (2017) 150:14–22. doi: 10.1016/j.neuroimage.2017.01.055

7. Chan MY, Park DC, Savalia NK, Petersen SE, Wig GS. Decreased segregation of brain systems across the healthy adult lifespan. *Proc Natl Acad Sci.* (2014) 111:E4997–5006. doi: 10.1073/pnas.1415122111

8. Vieira BH, Rondinoni C, Salmon C. Evidence of regional associations between age-related inter-individual differences in resting-state functional connectivity and cortical thinning revealed through a multi-level analysis. *Neuroimage.* (2020) 211:116662. doi: 10.1016/j.neuroimage.2020.116662

9. Finn ES, Shen X, Scheinost D, Rosenberg MD, Huang J, Chun MM, et al. Functional connectome fingerprinting: Identifying individuals using patterns of brain connectivity. *Nat Neurosci.* (2015) 18:1664. doi: 10.1038/nn.4135

10. Kaufmann T, Alnaes D, Doan NT, Brandt CL, Andreassen OA, Westlye LT. Delayed stabilization and individualization in connectome development are related to psychiatric disorders. *Nat Neurosci.* (2017) 20:513. doi: 10.1038/nn.4511

11. Knights E, Morcom AM, Henson RN. Does hemispheric asymmetry reduction in older adults (HAROLD) in motor cortex reflect compensation. *J Neurosci.* (2021) 41:9361–73. doi: 10.1523/JNEUROSCI.1111-21.2021

12. Lehmann BC, Henson RN, Geerligs L, Cam-Can, White SR. Characterising group-level brain connectivity: A framework using Bayesian exponential random graph models. *Neuroimage.* (2020) 225:117480. doi: 10.1016/j.neuroimage.2020.117480

13. Geerligs L, Tsvetanov KA, Cam-Can, Henson RN. Challenges in measuring individual differences in functional connectivity using fMRI: The case of healthy aging. *Hum Brain Mapp.* (2017) 38:4125–56. doi: 10.1002/hbm.23653

14. Shafto MA, Tyler LK, Dixon M, Taylor JR, Rowe JB, Cusack R, et al. The Cambridge centre for ageing and neuroscience (Cam-CAN) study protocol: A cross-sectional, lifespan, multidisciplinary examination of healthy cognitive ageing. *BMC Neurol.* (2014) 14:204. doi: 10.1186/s12883-014-0204-1

15. Yan C, Wang X, Zuo X, Zang Y. DPABI: Data processing & analysis for (resting-state) brain imaging. *Neuroinformatics.* (2016) 14:339–51. doi: 10.1007/s12021-016-9299-4

16. Ashburner J. A fast diffeomorphic image registration algorithm. *Neuroimage.* (2007) 38:95–113. doi: 10.1016/j.neuroimage.2007.07.007

17. Friston KJ, Williams S, Howard R, Frackowiak RS, Turner R. Movement-related effects in fMRI time-series. *Mag Reson Med.* (1996) 35:346–55. doi: 10.1002/mrm.1910350312

18. Fox MD, Zhang D, Snyder AZ, Raichle ME. The global signal and observed anticorrelated resting state brain networks. *J Neurophysiol.* (2009) 101:3270–83. doi: 10.1152/jn.90777.2008

19. Murphy K, Birn RM, Handwerker DA, Jones TB, Bandettini PA. The impact of global signal regression on resting state correlations: Are anti-correlated networks introduced? *Neuroimage.* (2009) 44:893–905. doi: 10.1016/j.neuroimage.2008.09.036

20. Murphy K, Fox MD. Towards a consensus regarding global signal regression for resting state functional connectivity MRI. *Neuroimage.* (2016) 154:169–73. doi: 10.1016/j.neuroimage.2016.11.052

21. Joliot M, Jobard G, Naveau M, Delcroix N, Petit L, Zago L, et al. AICHA: An atlas of intrinsic connectivity of homotopic areas. *J Neurosci Methods.* (2015) 254:46–59. doi: 10.1016/j.jneumeth.2015.07.013

22. Samper-González J, Burgos N, Bottani S, Fontanella S, Lu P, Marcoux A, et al. Reproducible evaluation of classification methods in Alzheimer's disease: Framework and application to MRI and PET data. *NeuroImage.* (2018) 183:504–21. doi: 10.1016/j.neuroimage.2018.08.042

23. Ezequiel G, Brent M, Sonal B, Vandergrift WA, Chris R, Carrie MD, et al. Deep learning applied to whole-brain connectome to determine seizure control after epilepsy surgery. *Epilepsia.* (2018) 59:1643–54. doi: 10.1111/epi.14528

24. Joliot M, Tzourio-Mazoyer N, Mazoyer B. Intra-hemispheric intrinsic connectivity asymmetry and its relationships with handedness and language Lateralization. *Neuropsychologia.* (2016) 93:437–47. doi: 10.1016/j.neuropsychologia.2016.03.013

25. Mesulam MM. *Principles of behavioral and cognitive neurology.* Oxford: Oxford university press (2000).

26. Luo N, Sui J, Abrol A, Turner JA, Damaraju E, Fu Z, et al. Structural brain architectures match intrinsic functional networks and vary across domains: A study from 15000+ individuals. *Cereb Cortex.* (2020) 30:5460–70. doi: 10.1093/cercor/bhaa127

27. Chang M, Womer FY, Edmiston EK, Bai C, Zhou Q, Jiang X, et al. Neurobiological commonalities and distinctions among three major psychiatric diagnostic categories: A structural MRI study. *Schizophr Bull.* (2018) 44:65–74. doi: 10.1093/schbul/sbx028

28. Anderson KM, Ge T, Kong R, Patrick LM, Spreng RN, Sabuncu MR, et al. Heritability of individualized cortical network topography. *Proc Natl Acad Sci.* (2021) 118:e2016271118. doi: 10.1073/pnas.2016271118

29. Geerligs L, Maurits NM, Renken RJ, Lorist MM. Reduced specificity of functional connectivity in the aging brain during task performance. *Hum Brain Mapp.* (2013) 35:319–30. doi: 10.1002/hbm.22175

30. Carp J, Park J, Polk TA, Park DC. Age differences in neural distinctiveness revealed by multi-voxel pattern analysis. *Neuroimage.* (2011) 56:736–43. doi: 10.1016/j.neuroimage.2010.04.267

31. Linda G, Renken RJ, Emi S, Maurits NM, Lorist MM. A brain-wide study of age-related changes in functional connectivity. *Cereb Cortex.* (2015) 25:1987–99. doi: 10.1093/cercor/bhu012

32. Patil AU, Madathil D, Huang CM. Healthy aging alters the functional connectivity of creative cognition in the default mode network and cerebellar network. *Front Aging Neurosci.* (2021) 13:607988. doi: 10.3389/fnagi.2021.607988

33. Fan L, Li H, Zhuo J, Zhang Y, Wang J, Chen L, et al. The human brainnetome atlas: A new brain atlas based on connectional architecture. *Cereb Cortex.* (2016) 26:3508–26. doi: 10.1093/cercor/bhw157

34. Dresler M, Shirer WR, Konrad BN, Müller NCJ, Wagner IC, Fernández G, et al. Mnemonic training reshapes brain networks to support superior memory. *Neuron.* (2017) 93:1227–35. doi: 10.1016/j.neuron.2017.02.003

35. Shi W, Fan L, Jiang T. Developing neuroimaging biomarker for brain diseases with a machine learning framework and the brainnetome atlas. *Neurosci Bull.* (2021) 37:1523–5. doi: 10.1007/s12264-021-00722-8

36. Zhu W, Huang H, Yang S, Luo X, Zhu W, Xu S, et al. Dysfunctional architecture underlies white matter hyperintensities with and without cognitive impairment. *J Alzheimers Dis.* (2019) 71:461–76. doi: 10.3233/JAD-190174

37. Zhu Y, Qi S, Zhang B, He D, Teng Y, Hu J, et al. Connectome-based biomarkers predict subclinical depression and identify abnormal brain connections with the lateral habenula and thalamus. *Front Psychiatry.* (2019) 10:371. doi: 10.3389/fpsyt.2019.00371

38. Hartwigsen G, Bengio Y, Bzdok D. How does hemispheric specialization contribute to human-defining cognition? *Neuron.* (2021) 109:2075–90. doi: 10.1016/j.neuron.2021.04.024

39. Tang P, Guo F, Xi Y, Peng L, Cui L, Wang H, et al. Distinct hemispheric specialization of functional connectivity in schizophrenia with and without auditory verbal hallucinations. *NeuroReport.* (2019) 30:1294–8. doi: 10.1097/WNR.0000000000001364

40. Long X, Jiang C, Zhang L. Morphological biomarker differentiating MCI converters from nonconverters: Longitudinal evidence based on hemispheric asymmetry. *Behav Neurol.* (2018) 2018:3954101. doi: 10.1155/2018/3954101

41. Johansson J, Salami A, Lundquist A, Wåhlin A, Andersson M, Nyberg L. Longitudinal evidence that reduced hemispheric encoding/retrieval asymmetry predicts episodic-memory impairment in aging. *Neuropsychologia.* (2019) 137:107329. doi: 10.1016/j.neuropsychologia.2019.107329

42. Lange N, Dubray MB, Lee JE, Froimowitz MP, Froehlich AF, Adluru N, et al. Atypical diffusion tensor hemispheric asymmetry in autism. *Autism Res.* (2011) 3:350–8. doi: 10.1002/aur.162

43. Ding Y, Yang R, Yan C, Chen X, Bai T, Bo Q, et al. Disrupted hemispheric connectivity specialization in patients with major depressive disorder: Evidence from the REST-meta-MDD Project. *J Affect Disord.* (2021) 284:217–28.

44. Mcginnis SM, Brickhouse M, Pascual B, Dickerson BC. Age-related changes in the thickness of cortical zones in humans. *Brain Topogr.* (2011) 24:279–91. doi: 10.1007/s10548-011-0198-6

45. Geerligs L, Cam-Can, Henson RN. Functional connectivity and structural covariance between regions of interest can be measured more accurately using multivariate distance correlation. *Neuroimage.* (2016) 135:16–31. doi: 10.1016/j.neuroimage.2016.04.047

46. Edelmann D, Mori TF, Szekely GJ. On relationships between the pearson and the distance correlation coefficients. *Stat Probab Lett.* (2021) 169:108960. doi: 10.1016/j.spl.2020.108960

TYPE Original Research
PUBLISHED 25 October 2022
DOI 10.3389/fpsyt.2022.1008011

OPEN ACCESS

EDITED BY
Xu Lei,
Southwest University, China

REVIEWED BY
Shenquan Liu,
South China University of Technology, China
Xianfa Jiao,
Hefei University of Technology, China

*CORRESPONDENCE
Xiaochuan Pan
pxc@ecust.edu.cn

SPECIALTY SECTION
This article was submitted to Computational Psychiatry, a section of the journal Frontiers in Psychiatry

RECEIVED 31 July 2022
ACCEPTED 10 October 2022
PUBLISHED 25 October 2022

CITATION
Zhang Y, Pan X and Wang Y (2022) Category learning in a recurrent neural network with reinforcement learning. *Front. Psychiatry* 13:1008011. doi: 10.3389/fpsyt.2022.1008011

Category learning in a recurrent neural network with reinforcement learning

Ying Zhang, Xiaochuan Pan* and Yihong Wang

Institute for Cognitive Neurodynamics, East China University of Science and Technology, Shanghai, China

It is known that humans and animals can learn and utilize category information quickly and efficiently to adapt to changing environments, and several brain areas are involved in learning and encoding category information. However, it is unclear that how the brain system learns and forms categorical representations from the view of neural circuits. In order to investigate this issue from the network level, we combine a recurrent neural network with reinforcement learning to construct a deep reinforcement learning model to demonstrate how the category is learned and represented in the network. The model consists of a policy network and a value network. The policy network is responsible for updating the policy to choose actions, while the value network is responsible for evaluating the action to predict rewards. The agent learns dynamically through the information interaction between the policy network and the value network. This model was trained to learn six stimulus-stimulus associative chains in a sequential paired-association task that was learned by the monkey. The simulated results demonstrated that our model was able to learn the stimulus-stimulus associative chains, and successfully reproduced the similar behavior of the monkey performing the same task. Two types of neurons were found in this model: one type primarily encoded identity information about individual stimuli; the other type mainly encoded category information of associated stimuli in one chain. The two types of activity-patterns were also observed in the primate prefrontal cortex after the monkey learned the same task. Furthermore, the ability of these two types of neurons to encode stimulus or category information was enhanced during this model was learning the task. Our results suggest that the neurons in the recurrent neural network have the ability to form categorical representations through deep reinforcement learning during learning stimulus-stimulus associations. It might provide a new approach for understanding neuronal mechanisms underlying how the prefrontal cortex learns and encodes category information.

KEYWORDS

category learning, stimulus-stimulus association, recurrent neural network, reinforcement learning, reward

Introduction

Category is a fundamental concept in cognitive neuroscience. The literature has demonstrated that humans and animals can use categorical information quickly and efficiently to identify new objects, make inference and so on (1–3). For example, we could classify an animal as a dog on the basis of its physical characteristics, even the animal would be a new type of dog that we did not know before. And we could infer its basic properties that belong to the dog category commonly. There are two types of category in the literature: perceptual category and functional category. Objects sharing similar physical properties could be classified into a group as a perceptual category (4). A functional category indicates that its members that share no any physical similarity have the similar function, such as associating the same action or reward (5–7), etc. Many behavioral studies suggest that animals could form a functional category of a group of visual stimuli through training the matching-to-sample task (8, 9). In this task, some arbitrarily selected visual images (samples) are learned to associate with a common target image. After learning, it is found that animals could treat these visual images as equivalent stimuli, known as a functional category (10, 11). It is an important research topic in the literature of studying the category that how animals or the neuronal system could learn, represent and utilize category information.

Various experimental data, including fMRI studies, lesion studies, and neurophysiological studies, demonstrated that rather than a single brain area, many brain areas are involved in the categorical processing, such as the inferior temporal cortex, the prefrontal cortex (PFC), and the basal ganglia (12, 13). Different brain areas may have distinct contributions toward processing category-related information. Neurons in the inferior temporal cortex are more sensitive to perceptual features of stimuli than categorical relations (14–16). Neurons in the PFC can achieve the categorical distinction based on abstract rules (17). PFC neurons have stronger category coding ability than do inferior temporal cortex neurons in categorization tasks (18, 19), and neurons show more similar responses to stimuli belonging to the same category than to stimuli belonging to different categories (20, 21). In addition, the execution of actions in categorization decision-making tasks requires not only the involvement of the premotor cortex but also relevant functions of the basal ganglia to help the PFC complete the adjustment of strategies. Thus, it has been reported that the premotor cortex and the basal ganglia are also engaged in category learning (22–25). Although it is known that many brain areas perform different functional roles during category learning, the mechanism underlying how these areas cooperate to learn and encode the category is unclear. Therefore, we try to construct a network model to further understand the working mechanism of the neural system in a categorization decision-making task. In particular, the PFC plays essential roles in processing category information and we build the network model to mimic functional roles of the PFC in the categorization decision-making task.

Some theoretical models have been proposed to explain how the category is learned in the neural system (26–28). But most of models show categorical phenomena that are consistent with some behavioral results, without showing neural activity that encodes category information observed in the PFC or other brain areas (29, 30). Hinaut and Dominey constructed a neural network model of the PFC that demonstrated how categorization of behavioral sequences can be achieved through a recurrent system (31). Their model is a three-layer cortical neural network that is sensitive to the sequence. As a result, a few neurons in the three-layer model could identify each sequence and a few other neurons produce an explicit representation of the category to which sequences belong. However, this neural network model is able to discriminate categories by using supervised learning, which is not biologically plausible for animals learning in the decision-making task. Experimental studies have demonstrated that animals learn to perform specific tasks based on the reward feedback for taking action (32), known as reinforcement learning (RL).

A large number of studies have shown that a combination of artificial neural networks with RL could make the network model learn and storage items more efficiently and faster (33, 34). In particular, the RL has been used to understand neural mechanisms of association learning in the cerebral cortex (35, 36). In the RL framework, the agent takes action by trial and error, and then it can obtain rewards from the external environment. Its purpose is to maximize the expected amount of reward (37). Surprisingly, the recurrent neural network trained with repeated RL can mimic the complex behavior of animals observed in various decision-making tasks (38, 39). However, in most of these studies, the recurrent network was trained to learn stimulus-action associations or stimulus-reward associations in the tasks with single decision-making. Few studies have reported that the recurrent neural network with RL could be applied in category learning. We are interested in whether this type of model could learn the functional category for a group of stimuli through stimulus-stimulus associations in the tasks with multiple decision-makings.

In this study, we constructed a deep RL model that combines a recurrent neural network with RL to investigate how the category is learned in the network. On the one hand, this network model uses the gated recurrent unit network structure where neurons can regulate information transmission through gating mechanisms. On the other hand, this network model utilizes the actor-critic algorithm structure where neurons can update weights and biases through the policy gradient RL algorithm (40). Then, we investigate whether this model can mimic the behavior of monkeys and their neural activities in the PFC reported in a sequential paired-association task (41).

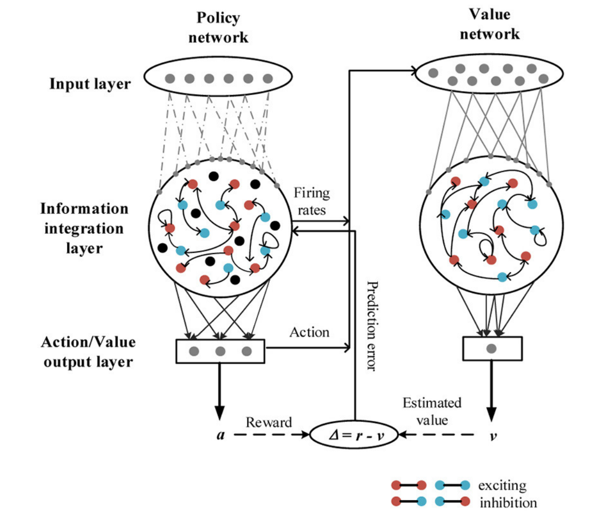


FIGURE 1
Structures of the neural network model. The deep RL neural network model, consisting of a policy network and a value network. In the policy network, sparse connections are made from the input layer to the information integration layer (IIL), among neurons in the IIL. Full connections are made from the IIL to the action output layer. In the value network, full connections are made among neurons in the input layer, the IIL, and the value output layer. In addition, in the IIL, red-red or blue-blue indicates excitatory connections between neurons; red-blue or blue-red indicates inhibitory connections between neurons; and black indicates no connection between neurons.

TABLE 1 Training parameters of the deep RL model.

Parameter	Value	Description
α	0.01	Learning rate
Δt	20 ms	Time step
τ	100 ms	Time constant
N_{p_in}	11	Number of neurons in the input layer (policy network)
N_{v_in}	153	Number of neurons in the input layer (value network)
N_p	150	Number of neurons in the IIL (policy network)
N_v	100	Number of neurons in the IIL (value network)
N_{p_out}	3	Number of neurons in the action output layer (policy network)
N_{v_out}	1	Number of neurons in the value output layer (value network)
p_0	0.2	Connection probability (policy network)
p_1	0.1	Connection probability (policy network)
p_2	1	Connection probability (policy network)
δ^2_{rec}	0.01	Network noise
N_{trials}	24	Number of trials for gradient update
T	121	Maximum time of per trial

In the sequential paired-association task, this model needs to learn six stimulus-stimulus associative sequences in a similar way to train the monkey to learn this task. It was found that the model was able to successfully learn the six associative sequences at the end of the training, reproducing the choice behavior of the monkey observed in the task. Notably, we found two types of neurons in this model: one type primarily encodes information about individual stimuli; the other type mainly encodes category information of associated stimuli in one chain. The ability of these two types of neurons to encode information was enhanced during the learning process of this model. Our results suggest that the neurons in the recurrent neural network have the ability to form categorical representations through deep RL during learning stimulus-stimulus associations.

Methods

Neural network model

The deep RL network has been used to simulate stimulus-response associations or stimulus-reward associations in previous studies (38, 42). In this study, a new neural network based on the framework of the deep RL is proposed. The deep RL neural network model is composed of two parts: the policy network and the value network (Figure 1).

The policy network has three layers: the input layer, the information integration layer (IIL), and the action output layer. The number of neurons in the input layer is $N_{p\_in} = 11$, and these neurons receive stimulus information from the external environment; the number of neurons in the IIL is $N_p = 150$, and these neurons can receive stimulus information from the input layer; the number of neurons in the action output layer is $N_{p\_out} = 3$, and these neurons represent three actions: fixation, left and right choices in this study. The probability of connection from each neuron in the input layer to neurons in the IIL is $p_0 = 0.2$; the probability of connection among neurons in the IIL is $p_1 = 0.1$; the probability of connection from each neuron in the IIL to neurons in the action output layer is $p_2 = 1$ (fully connected, see Table 1).

The value network also has three layers. The number of neurons in the input layer is $N_{v\_in} = 153$, and these neurons receive the firing rates of 150 neurons in the IIL and the action of 3 neurons in the action output layer of the policy network; the number of neurons in the IIL is $N_v = 100$, and these neurons can learn information from the policy network; the number of neurons in the value output layer is $N_{v\_out} = 1$, and the neuron generates a predictive reward for each action. Here, full connections are made among neurons in the input layer, the IIL and the value output layer.

In this model, the policy network generates an action based on current stimulus and task conditions, and this model takes the action and receives an actual reward; the value network integrates neuronal firing rates in the policy network to output a predictive reward for the action. There is a reward prediction error between the actual reward and the predictive reward for the action, and the policy network adjusts the policy in time according to the error signal to minimize it.

In both the policy network and value network, the IILs have a recurrent connection structure with gated recurrent units (a gated recurrent unit is considered as a neuron). The gated recurrent unit includes an update gate and a reset gate, where the update gate is used to control the retained historical state information and receives new information about the candidate state, and the reset gate is used to control the dependence on historical state information for candidate information (43). In this way, information forms a dependency between different states of the transmission process. In this paper, the equations of the continuous-time gated recurrent unit network for the policy network are described in Equations (1)–(4), and the value network has similar equations for its gated units.

$$\begin{aligned}
\phi_i(t) &= \sigma\left(\sum_{j=1}^{N_p} W_{rec}^{\phi,ji} x_j(t-1)\right. \\
&\quad \left. + \sum_{k=1}^{N_{p\_in}} W_{in}^{\phi,ki} u_k(t) + b_i^{\phi}(t)\right), \left(i = 1, \ldots, N_p\right), \qquad (1) \\
\psi_i(t) &= \sigma\left(\sum_{j=1}^{N_p} W_{rec}^{\psi,ji} x_j(t-1)\right. \\
&\quad \left. + \sum_{k=1}^{N_{p\_in}} W_{in}^{\psi,ki} u_k(t) + b_i^{\psi}(t)\right), \qquad (2) \\
h_i(t) &= \left(1 - \eta\phi_i(t)\right) h_i(t-1) \\
&\quad + \eta\phi_i(t)\left[\sum_{j=1}^{N_p} W_{rec}^{ji}\left(\psi_j(t)\, x_j(t-1)\right) + \sum_{k=1}^{N_{p\_in}} W_{in}^{ki} u_k(t)\right. \\
&\quad \left. + b_i(t) + \sqrt{2\eta^{-1}\delta_{rec}^2 \varepsilon}\right], \qquad (3) \\
x_i(t) &= \left[h_i(t)\right]^{+}. \qquad (4)
\end{aligned}$$

Here, we use the modified linear activation function $[x]^{+} = \max(0, x)$ as the output function of each neuron. Because the gated unit in GRU network is considered as the firing rate neuron, the value of its output function is defined as the firing rate of the neuron. The firing rate of each neuron in the IIL is non-negative. In addition, $\sigma(x) = \left[1 + \exp(-x)\right]^{-1}$ as the output function of each gate [the update gate $\phi_i(t)$ or the reset gate $\psi_i(t)$, $(i = 1, \ldots, N_p)$, $(t = 1, \ldots, T)$], ε is the Gaussian white noise with a mean of 0 and variance of 1, and δ_{rec}^2 is used to control the size of this network noise. And $u_k(t)$ $(k = 1, \ldots, N_{p\_in})$ is the input information of the *kth* neuron from the external environment at time t, $x_i(t)$ is the firing rate of the *ith* neuron at time t. $\eta = \Delta\frac{t}{\tau}$, Δt is the time step, and τ is the time constant (Table 1), which is used to control the information dependency of gate recurrent units. $W_{rec}^{\phi,ji}$ and $W_{rec}^{\psi,ji}$ are the connection weights from the *jth* neuron to the *ith* neuron in the update gate and reset gate (44), respectively; $W_{in}^{\phi,ki}$ and $W_{in}^{\psi,ki}$ are the connection weights from the *kth* input neuron to the *ith* neuron in the update gate and reset gate, respectively; $b_i^{\phi}(t)$ and $b_i^{\psi}(t)$ are the bias of the update gate and reset gate, respectively. In addition, W_{rec}^{ji} is the connection weight from the *jth* neuron to the *ith* neuron in the IIL; W_{in}^{ki} is the connection weight from the *kth* neuron in the input layer to the *ith* neuron in the IIL; $b_i(t)$ is the bias of the *ith* neuron in the IIL.

Specifically, $x_i^{\pi}(t)$ is the firing rate of the *ith* neuron in the IIL of the policy network under the policy of π. Generally speaking, RL consists of five main elements: an agent, an environment, actions, states, and rewards. The agent first observes the external environment and receives the input information u_t (the N_{p_in} dimensional vector), and then according to the policy $\pi_\theta(a_t|u_t)$ chooses an action a_t (the N_{p_out} dimensional vector). Here, the action output layer neurons generate an action based on the policy function:

$$\begin{aligned}
z_l(t) &= \sum_{i=1}^{N_p} W_{out}^{\pi,il} x_i^{\pi}(t) \\
&\quad + b_{out}^{\pi,l}(t), \left(l = 1, \ldots, N_{p-out}\right), \qquad (5) \\
\pi_\theta\left(a_t = l|u_t\right) &= \frac{e^{z_l(t)}}{\sum_{l=1}^{N_{Pout}} e^{z_l(t)}}. \qquad (6)
\end{aligned}$$

Where $W_{out}^{\pi,il}$ $(l = 1, \ldots, N_{p-out})$ is the connection weight from the *ith* neuron in the IIL to the *lth* neuron in the action output layer of the policy network, $b_{out}^{\pi,l}(t)$ is the bias of the *lth* neuron in the action output layer, $z_l(t)$ is the linear output function of the *lth* neuron in the action output layer, and the policy $\pi_\theta(a_t|u_t)$ is the softmax function. The agent chooses an action based on the policy function through the random sampling method. That is to say, when the agent has very limited information about the external environment from observation, it cannot completely rely on the information to make a correct choice. However, the agent obtains a reward provided by the environment in the occasional event of taking action. In this case, an evaluation of the action by the value network can better guide the policy network to implement the adjustment of the policy. Here, the firing rate of the *mth* neuron in the IIL of the value network is $x_m^{v}(t)$ $(m = 1, \ldots, N_v)$, and the neuron in the value output layer generates a predictive reward for the action based on the value function:

$$v_\varphi\left(x_t^{\pi}, a_t\right) = \sum_{m=1}^{N_v} W_{out}^{v,m} x_m^{v}(t) + b_{out}^{v}(t). \qquad (7)$$

Where the firing rate x_t^{π} (the N_p dimensional vector) of neurons in the IIL of the policy network and the action a_t (the N_{p_out} dimensional vector) of neurons in the action output layer as the input information of the value network. $W_{out}^{v,m}$ is the connection weight from the *mth* neuron in the IIL to the neuron in the value output layer, $b_{out}^{v}(t)$ is the bias of the neuron in the value output layer, and v_{φ} is the linear output information of the value output layer.

Policy gradient reinforcement learning

In this model, the connection weights (W_{in}, W_{in}^{ϕ}, W_{in}^{ψ}, W_{rec}, W_{rec}^{ϕ}, W_{rec}^{ψ}, W_{out}^{π}, and W_{out}^{v}) and biases (b, b^{ϕ}, b^{ψ}, b_{out}^{π}, and b_{out}^{v}) of neurons are updated by the policy gradient RL algorithm during training (38). In this study, considering that the environmental state for the agent is not completely observable, we use a partially observable Markov decision process model, which is more suitable for the agent to learn in the state of limited information about the external environment. The partially observable Markov decision process model is discrete and finite (45). The continuous period is discretized through time steps, and the agent needs to observe the external environment and chooses an action at every time step. Setting the time ranges from 0 to time t, $I_{0:t}$ is the historical information in the interaction process between the agent and the environment, including the states, observations, and actions, as follows:

$$I_{0:t} = \left(s_{0:t+1}, u_{1:t}, a_{0:t}\right). \tag{8}$$

After the agent chooses an action a_t at the time t, it obtains a reward r_{t+1} at the next time $t + 1$. In detail, when $t = 0$, the environment is in the current state s_0 with the probability $\kappa(s_0)$, and the agent chooses an action a_0 according to the policy π_{θ}, where θ denotes the parameter, including the weights and biases of the policy network. When $t = 1$, the environment enters the new state s_1 with the probability $\kappa\left(s_1|s_0, a_0\right)$, and the agent obtains a reward r_1. Next, the agent observes the external environment and receives the input u_1, and chooses an action a_1 based on the new policy $\pi_{\theta}(a_1|u_1)$ and obtains a reward r_2. Thus, a process of the interaction between the agent and the environment is to keep repeating these steps until the end of each trial. In general, from the beginning to the end of each trial, the agent can rely on the policy π_{θ} at time t to choose an action a_t that eventually obtain the maximum expected value of the reward $R(\theta)$:

$$R(\theta) = E_I\left(\sum_{t=0}^{T} r_{t+1}\right). \tag{9}$$

Where the T is the end time of each trial (Table 1), and the E_I is the expected calculation on the basis of the history $I_{0:T} = \left(s_{0:T+1}, u_{1:T}, a_{0:T}\right)$.

Our model utilizes the policy gradient method with an actor-critic algorithm structure when updating parameters. This approach uses the policy function and the value function for learning. Briefly, the actor takes action by adjusting the policy, which is the policy function; the critic evaluates each policy by predicting the reward of this action, known as the value function.

In order to update parameters of the policy network (actor) by the gradient descent method, an objective function is defined as follows:

$$\Gamma^{\pi}(\theta) = \frac{1}{N_{trials}} \sum_{n=1}^{N_{trials}} -R_n(\theta). \tag{10}$$

Where the parameter θ includes the weights and biases of the policy network. Notably, when training the network model, we did not update parameters of the policy network in every trial; instead updating those after the completion of N_{trials} trials. This method makes learning process of the policy network more stable. In addition, we use the policy gradient algorithm to solve $\nabla_{\theta} R_n(\theta)$:

$$\nabla_{\theta} R_n(\theta) = \sum_{t=0}^{T} \nabla_{\theta} \log \pi_{\theta}(a_t|u_t)\, \Upsilon\left(x_t^{\pi}, a_t\right), \tag{11}$$

$$\Upsilon\left(x_t^{\pi}, a_t\right) = \sum_{t=0}^{T} r_{t+1} - v_{\varphi}\left(x_t^{\pi}, a_t\right). \tag{12}$$

Here, the $\Upsilon\left(x_t^{\pi}, a_t\right)$ is a reward prediction error value of the Temporal-Difference algorithm, which denotes the difference between the estimated value of the value function and the actual reward. This value can be used as an error signal to guide the policy network to learn. At the time t, $v_{\varphi}\left(x_t^{\pi}, a_t\right)$ is the linear output function of the value network, and x_t^{π} is the firing rates of neurons in the IIL of the policy network.

In order to update parameters of the value network (critic) by the gradient descent method, an objective function is defined as follows:

$$\Gamma^{v}(\varphi) = \frac{1}{N_{trials}} \sum_{n=1}^{N_{trials}} M_n(\varphi), \tag{13}$$

$$M_n(\varphi) = \frac{1}{T+1} \sum_{t=0}^{T} \left[r_{t+1} - v_{\varphi}\left(x_t^{\pi}, a_t\right)\right]^2. \tag{14}$$

Where $M_n(\varphi)$ is the mean square error, and the parameter φ includes the weights and biases of the value network. In the value network, the firing rates x_t^{π} of neurons in the IIL of the policy network and the action a_t of neurons in the action output layer as its input information at time t, and its output information is a prediction value v_{φ} of the action. Here, we solve $\nabla_{\varphi} M_n(\varphi)$ by Backpropagation through the time algorithm (46). Finally, our model can learn dynamically based on the interaction of information between the policy network and the value network.

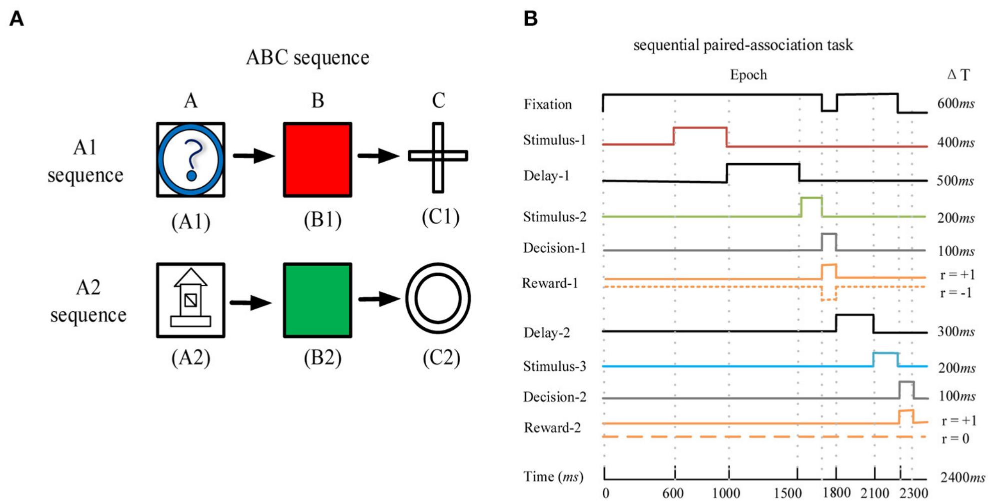


FIGURE 2
The sequential paired-association task and its task events. **(A)** The example of the ABC sequence learned by the monkey. The two correct stimulus-stimulus associative sequences are A1→ B1→ C1 and A2→ B2→ C2. **(B)** Timing of task events in a trial of the sequential paired-association task. The network model needs to fixate on the fixation spot during the stimulus and delay periods. It obtains a positive reward $r_{t+1} = +1$ for each correct choice during the two decision periods (Decision-1 and Decision-2). If this model makes a wrong choice in the first decision period, it will obtain a negative reward $r_{t+1} = -1$ and the current trial is terminated. If this model makes a wrong choice in the second decision period, it will not obtain a reward ($r_{t+1} = 0$) and the trial will end.

Sequential paired-association task

We used the deep RL model to learn the sequential paired-association task that has been performed successfully by the monkey (41). In this task, the monkey needed to learn two stimulus-stimulus associative sequences (Figure 2A). Here, the visual stimuli were six distinguishable pictures, which were divided into two associative sequences (A1→B1→C1 and A2→B2→C2). Figure 2B shows task events that are suitable for this model to learn. The maximum time of each trial is 2,400 ms (Figure 2B). At the beginning of each trial, the agent is required to fixate on the fixation spot for 600 ms. After that, the first stimulus A1 or A2 is presented for 400 ms. Following the first stimulus, there is a delay period of 500 ms. The agent continues fixating on the spot during the delay period. After the delay, the second stimuli B1 and B2 are presented simultaneously on the left and right positions. The left and right positions of the two stimuli are random. At this time, the agent is required to fixate on the spot for 200 ms. After the second stimuli is offset, the agent is given 100 ms to make the first choice (selection of B1 or B2 based on A1 or A2). If the first choice is wrong and the current trial is terminated. If the first choice is correct, the agent obtains a reward and the trial is to be continued. After the first correct choice, the agent is required to fixate on the spot for 300 ms. Then the third stimuli C1 and C2 are presented simultaneously, and the left and right positions of the two stimuli are random. At this time, the agent is required to fixate on the spot for 200 ms. After the third stimuli is offset, the agent is given 100 ms to make the second choice (selection of C1 or C2 based on B1 or B2). When the second choice is correct, the agent obtains a reward again and the trial is to end. The design of two associative sequences (A1-sequence and A2-sequence) allows the network model to select the target stimuli from the presentation of the target and distractor stimuli.

In the policy network, 11 neurons in the input layer denote the fixation, stimulus A1, stimulus A2, left stimulus B1, right stimulus B1, left stimulus B2, right stimulus B2, left stimulus C1, right stimulus C1, left stimulus C2, and right stimulus C2, respectively. In the sequential paired-association task, the fixation is labeled as a value of 1, the stimulus A1 or A2 is labeled as a value of 2, the stimulus B1 or B2 is labeled as a value of 3, and the stimulus C1 or C2 is labeled as a value of 4. The agent needs to take three actions ($N_{out} = 3$), and the three neurons in the action output layer are fixation ($a_t = F$), left ($a_t = L$), and right($a_t = R$), respectively. We choose appropriate values for the number of neurons in the two IILs ($N = 150$ in the policy network and $N = 100$ in the value network) and their connection probabilities (see Table 1) in order to enable the model to learn the task successfully. We did not systemically analyze how changes of these super-parameters affect the model to learn the task. However, the combination of appropriate values of these super-parameters is important for the model to learn the task.

In general, the agent can choose left or right action only during two decision periods; and it must keep fixation during the stimulus period and the delay period. When the agent chooses a correct action in the first decision period, it obtains a positive reward $r_{t+1} = +1$; when the agent chooses a wrong action in the first decision period, it obtains a negative reward $r_{t+1} = -1$ and the trial is terminated. The agent obtains a positive reward $r_{t+1} = +1$ for the correct action or a reward $r_{t+1} = 0$ for the wrong action in the second decision period. If the agent does not make a choice (left or right) during the second decision period, it obtains a negative reward $r_{t+1} = -1$. During the stimulus period or the delay period, the agent chooses the fixation action to receive a reward $r_{t+1} = 0$; if the agent chooses a left or right action, it obtains a negative reward $r_{t+1} = -1$ and the trial is terminated.

The model is required to learn not only the ABC sequence (A1→ B1→ C1 and A2→ B2→ C2), but also the BCA sequence (B1→ C1→ A1 and B2→ C2→ A2) and the CAB sequence (C1→ A1→ B1 and C2→ A2→ B2). The three sequences have similar task events in a trial. We divided the six stimuli A1, A2, B1, B2, C1, and C2 into two groups, the A1-group (A1, B1, and C1) and the A2-group (A2, B2, and C2). The stimuli in the A1-group are associated each other in one chain and the stimuli in the A2-group are associated each other in another chain. When this model is trained, the three sequences (ABC, BCA, and CAB) appear randomly in the learning process, and the agent learns six stimulus-stimulus associative sequences in parallel.

In this task, we set the time constant τ to 100 ms, the time step Δt to 20 ms, and the number of trials N_{trials} to 24, which denotes this network model updating parameters after 24 trials are completed (labeled as one iteration). In addition, when the network model completes 50 policy iterations, we test the network model with 800 trials to determine whether the policy is optimal. During the training process, the network model goes through the learning stage and testing stage alternately. The agent updates parameters through policy iterations in the learning stage, and the agent evaluates each policy without updating parameters in the testing stage. When the correct rate of choice (the ratio of correct trials to all trials) reaches 98% in the testing stage, we consider that the agent has found the optimal policy, which indicates that the network model can complete the task successfully.

The sequential paired-association task does not require the monkey to encode category information for the associated stimuli. Behaviorally, just memorizing each stimulus-stimulus association is sufficient for the monkey to perform the task successfully. However, it was reported that some prefrontal neurons encoded category information for the associated stimuli after the monkey learned the task (41, 47). We are interested in whether and how the network model forms categorical representations for associated stimuli during its learning of the sequential paired-association task.

Category index and stimulus index

After this model learned stimulus-stimulus associations, we further examined the activity of 150 neurons in the IIL of the policy network. To characterize the response of each neuron, we calculate the category index for each of them during the first stimulus period (0–400 ms from the first stimulus onset). First, for each neuron, we calculate the absolute value of the firing rate difference of every two stimuli from the A1-group, which is denoted as FD_{A1}. Similarly, we calculate the absolute value of the firing rate difference of every two stimuli from the A2-group, which is denoted by FD_{A2}. Then, we calculate the mean firing rate difference of stimuli within a category for each neuron, which is denoted by *WCD*. The equations are as follows:

$$WCD = \frac{FD_{A1} + FD_{A2}}{6}, \tag{15}$$

$$FD_{A1} = |x_{A1} - x_{B1}| + |x_{A1} - x_{C1}| + |x_{B1} - x_{C1}|, \tag{16}$$

$$FD_{A2} = |x_{A2} - x_{B2}| + |x_{A2} - x_{C2}| + |x_{B2} - x_{C2}|. \tag{17}$$

Where || indicates the absolute value. x_{A1}, x_{B1}, and x_{C1} denote the firing rate of each neuron to stimuli in the A1-group during the first stimulus period; x_{A2}, x_{B2}, and x_{C2} denote the firing rate of each neuron to stimuli in the A2-group during the first stimulus period. After that, we also calculate the absolute value of the firing rate difference of each neuron between every two stimuli across the two groups. Thus, the difference value between two categories is denoted by *BCD*. The equations are as follows:

$$BCD = \frac{FD_1 + FD_2 + FD_3}{9}, \tag{18}$$

$$FD_1 = |x_{A1} - x_{A2}| + |x_{A1} - x_{B2}| + |x_{A1} - x_{C2}|, \tag{19}$$

$$FD_2 = |x_{B1} - x_{A2}| + |x_{B1} - x_{B2}| + |x_{B1} - x_{C2}|, \tag{20}$$

$$FD_3 = |x_{C1} - x_{A2}| + |x_{C1} - x_{B2}| + |x_{C1} - x_{C2}|. \tag{21}$$

Finally, we define the category index according to *WCD* and *BCD*, which is denoted by *CI*, and it is given by:

$$CI = \frac{BCD - WCD}{BCD + WCD}. \tag{22}$$

The range of *CI* is from −1 to 1. When the category index is negative, the neuron shows larger response-differences to stimuli within a category than to stimuli across the two categories. When the category index is positive, the neuron shows larger response-differences to stimuli across the two categories than to stimuli within a category.

Bootstrap test is used to determine whether the category index of each neuron is statistically significant from zero or not. We shuffled its firing rates among the six stimuli (A1, B1, C1, A2, B2, and C2) in the first stimulus period and calculated the category index based on the shuffled data. This process was repeated 500 times, generating a distribution of shuffled category indexes. The original category index value was deemed significant if it fell within the top or bottom 2.5% of the shuffled distribution ($p < 0.05$).

In addition, noting that some neurons show differential activity to stimuli from a category, we define the stimulus index for each neuron based on its firing rates to the three stimuli in the same category during the first stimulus period (48), denoted by *SI*, which is calculated as follows:

$$SI = \frac{SI_{A1} + SI_{A2}}{2}, \tag{23}$$

$$SI_{A1} = \frac{x(A1)_{\max} - x(A1)_{\min}}{x(A1)_{\max} + x(A1)_{\min}}, \tag{24}$$

$$SI_{A2} = \frac{x(A2)_{\max} - x(A2)_{\min}}{x(A2)_{\max} + x(A2)_{\min}}. \tag{25}$$

Where $x(A1)_{\max}$ denotes the maximum firing rate of each neuron to the three stimuli (A1, B1, and C1) in the A1-group during the first stimulus period, and $x(A1)_{\min}$ denotes the minimum firing rate to the three stimuli. $x(A2)_{\max}$ denotes the maximum firing rate of each neuron to the three stimuli in the A2-group during the first stimulus period, and $x(A2)_{\min}$ denotes the minimum firing rate to the three stimuli. The *SI* reflects response-differences to stimuli within a category, ignoring response-differences to stimuli across the categories. The range of *SI* is from 0 to 1, $SI = 0$ indicates that the neuron shows no differential activity to stimuli from a category, but it may have differential activity to stimuli from different categories.

Results

Our model was performed using theano0.8.2 based on Python2.7 software in Windows 10 system, and the model was able to run successfully in learning the sequential paired-association task.

Behavior performance of the network model

The model was trained to learn the six stimulus-stimulus associations in parallel. In each trial, one of the six associations was inputted into the model. After 500 policy iterations, the network model could achieve the correct rate (the ratio of correct trials to all trials) of 98% in the two decision periods, indicating that it learned the sequential paired-association task (Figures 3A,B). It was worth noting that our network model needed to make two choices in each trial. In the early learning stage, the network model was trained to improve the correct rate of the first choice, the correct rate of the second choice was low. For example, the correct rate of the first choice and second choice were about 1.8 and 0% at the 50th policy iteration, respectively. When the network model increased gradually the correct rate of the first choice, it started to increase the correct rate of the second choice. At the 200th policy iteration, the correct rate of the first choice was about 25.4% and the correct rate of the second choice was about 12.6%. We found that from the 200th policy iteration, the mean square error (MSE) of reward prediction for the network model at the second choice decreased gradually during the training process (Figure 3C). It indicated that the predictive reward for the action estimated by the value network was getting closer to the actual reward. The result reflected that the network model could adjust the policy and choose a correct action in time through the feedback information provided by the error signal. The results suggested that our model could learn that the sequential paired-association task in different learning stages. Finally, this model was able to get the maximum reward in each trial (Figure 3D). The trained network model could reproduce the similar behavior of the monkey in the sequential paired-association task (41). It demonstrated that the model was able to learn stimulus-stimulus associative sequences.

Various activity-patterns of neurons

The output actions of this model demonstrated that it was able to correctly choose a target stimulus on the basis of the sample stimuli, indicating the model remembered stimulus-stimulus relations. How did neurons encode stimulus information and stimulus-stimulus relations to make a choice in our model? To investigate this issue, we further analyzed activity-patterns of neurons in the IIL of the policy network. Interestingly, neurons could produce various types of activity-patterns after our model learned the sequential paired-association task. During the first stimulus period (from 0 to 400 ms after the first stimulus onset), some neurons showed different responses to stimuli in the A1-group and the A2-group. For example, there are 19 neurons (19/150; 12.7%) produced excitatory activity to stimuli in the A1-group, and less activity to stimuli in the A2-group compared with the baseline activity (−200 to 0 ms from the first stimulus onset) (Figure 4A). Some neurons produced excitatory activity to stimuli in the A2-group and less activity to stimuli in the A1-group (Figure 4B), and the number of this type of neurons is 27 (27/150; 18%). About 14% (21/150) neurons produced excitatory activity to stimuli in the both A1-group and A2-group compared to the baseline activity (Figures 4C,E,F). In contrast, about 14.7% (22/150) neurons produced inhibitory activity to stimuli in the both A1-group and A2-group (Figure 4D). We also found that 16 neurons (16/150; 10.7%) showed no differential activity to stimuli in the both A1-group and A2-group (Figure 4G). Finally, about one third of neurons (45/150; 30%) kept silent during the whole trial (the firing rate of neurons was zero) (Figure 4H).

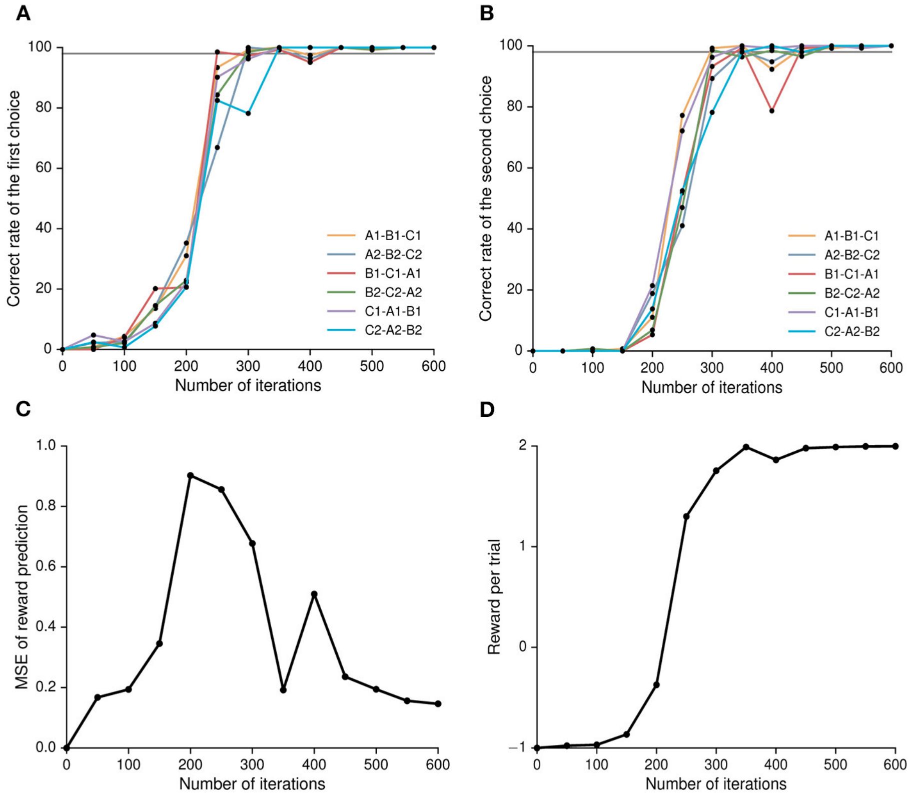


FIGURE 3
Behavior performance of the deep RL model. **(A)** Correct rate of the first decision period (the ratio of correct choice trials in the first decision period to all trial) for each stimulus-stimulus associations. **(B)** Correct rate of the second decision period (the ratio of correct choice trials in both decision periods to all trials) for each stimulus-stimulus associations. Here, the gray line denotes 98% of the target value. **(C)** The mean square error (MSE) of reward prediction for the network model in the second decision period (see Equation 14). Mean square error between the actual reward (based on the selected action in the policy network) and the predictive reward (estimated in the value network). **(D)** The reward obtained by the network model per trial.

Stimulus-neurons and category-neurons

Neurons in the IIL showed various types of activity-patterns. One important question is what kind of information these neurons encode in the model. We found that some neurons produced similar activity-patterns to the stimuli belonged to the same group, and differential activity-patterns to the stimuli belonged to different groups (see Figures 4A–D). The activity-patterns of these neurons were similar to those of PFC neurons observed in the sequential paired-association task (41). Many studies have demonstrated that PFC neurons can encode the category to which visual stimuli belong (49, 50). We hypothesized that neurons in this model could encode category information for each group of stimuli during stimulus-stimulus association learning.

To demonstrate whether the neuron in our model was able to represent categorical information, we calculated the category index for each neuron in the first stimulus period. According to the definition of category index (see Section Methods), we calculated the category indexes of 105 neurons (excluding 45 no-response neurons shown in Figure 4H), and the range is from −0.2 to 1 (Figure 5A). We noted that some neurons had negative category indexes, indicating these neurons encode less category information, whereas some neurons had positive category indexes encoded more category information. In order to determine whether the category index of individual neuron is significantly different from zero, the bootstrap method was used (see Section Methods). The results showed that 58 neurons in this IIL had an insignificant category index ($p > 0.05$) and the mean category index of these neurons was 0.243. We thought that these neurons could not identify the category to which the stimulus belongs, but encoded stimulus identity. These neurons are referred to as stimulus-neurons. In addition, 47 neurons had a significant category index ($p < 0.05$) and the mean category index of these neurons was 0.731. These neurons primarily encoded category information, denoted as category-neurons. It suggested that there were individual neurons having the ability to encode category information in our model.

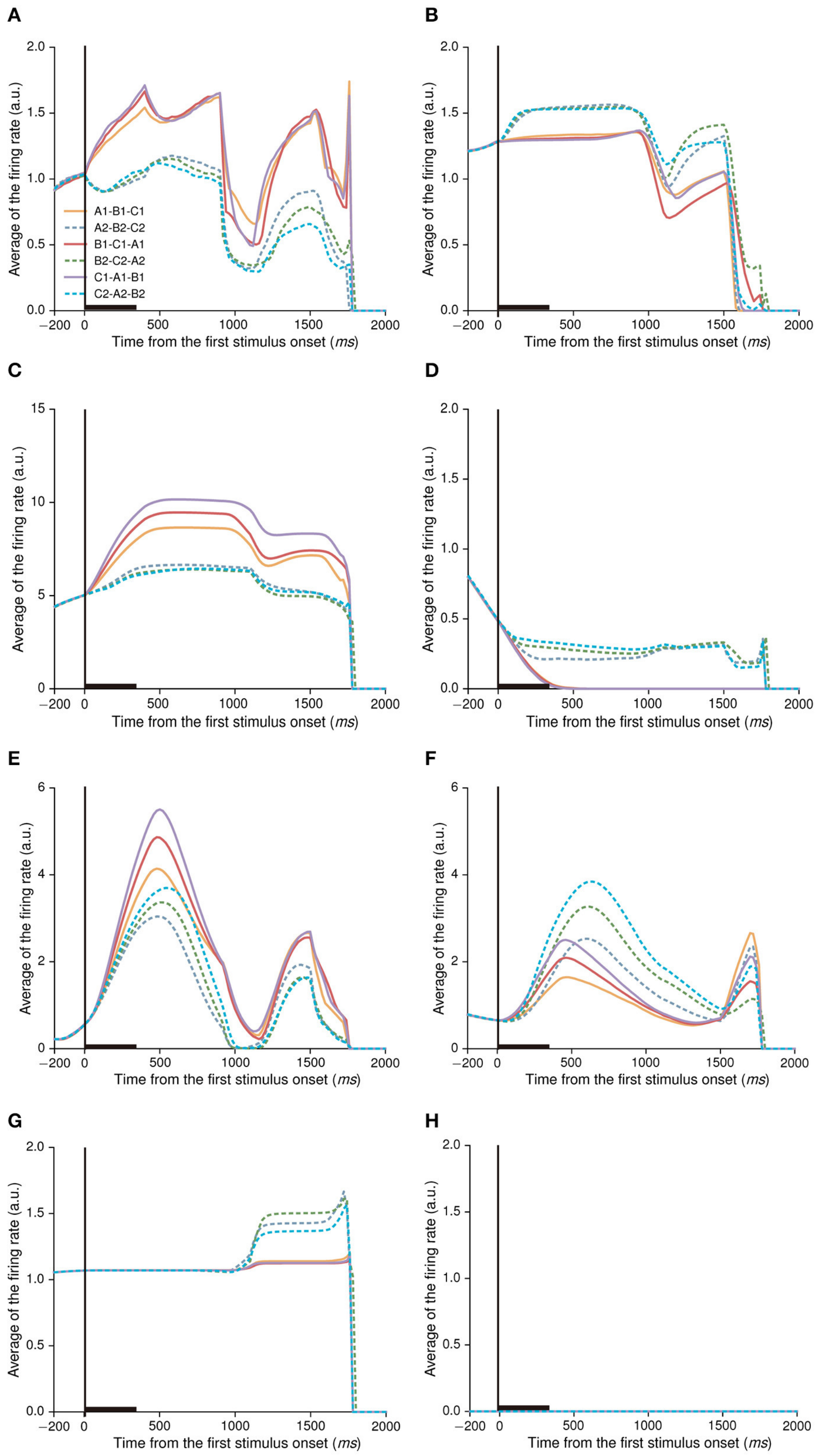


FIGURE 4

Various types of activity-patterns found in the IIL of the policy network after the model learned the task. Here, the black rectangle on the horizontal axis denotes the first stimulus period (0–400 ms from the first stimulus onset). During the first stimulus period, neurons show different

(Continued)

FIGURE 4 (Continued)
activity in response to stimulus A1, A2, B1, B2, C1, and C2. Neural activity is sorted by the six stimulus-stimulus associations. If the firing rate in the first stimulus period increases compared to the baseline activity (a period of−200 ms-0 from the first stimulus onset), it indicates an excitatory response. If the firing rate decreases, it indicates an inhibitory response. The averaged firing rate of one neuron indicates that its firing rates are averaged across all trials. The same figure legends are used in **(A–G)**. **(A–D)** show example activity of category-neurons ($p < 0.05$, Bootstrap test), the category indexes of these neurons are 0.820, 0.872, 0.619, and 0.627, respectively. **(E–G)** show example activity of stimulus-neurons ($p > 0.05$, Bootstrap test), the category indexes of these neurons are 0.393, −0.135, and 0.416, respectively. **(H)** An example neuron showing no activity in the whole trial.

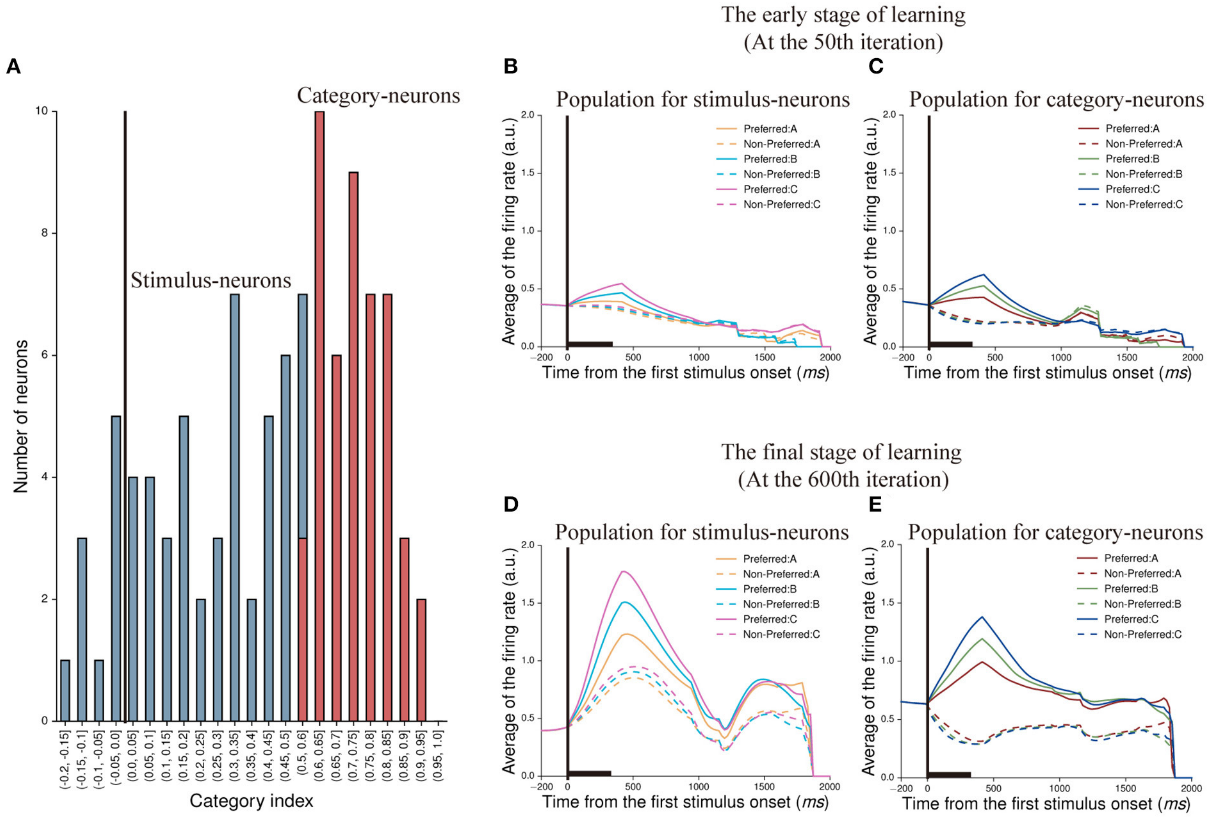


FIGURE 5
Classification of neurons and their population activity at two learning stages. **(A)** The distribution of category indexes. Here, blue bars indicate 58 neurons whose category indexes are not significant ($p > 0.05$, Bootstrap test), denoted as stimulus-neurons. And the range of category indexes for these neurons is from −0.2 to 0.6. Red bars indicate 47 neurons whose category indexes are significant ($p < 0.05$, Bootstrap test), denoted as category-neurons. And the range of category indexes for these neurons is from 0.5 to 1. **(B,C)** show population activity of stimulus-neurons **(B)** and category-neurons **(C)** in the early stage of learning (at the 50th iteration), respectively. The activity of each neuron is sorted by its preferred activity to the three paired stimuli (A1 vs. A2, B1 vs. B2, and C1 vs. C2) and then was averaged across neurons. **(D,E)** show population activity of stimulus-neurons **(D)** and category-neurons **(E)** in the final stage of learning (at the 600th iteration). The averaged firing rates shown in **(B,D)** are firing rates averaged across trials and across the stimulus-neurons. The averaged firing rates shown in **(C,E)** are firing rates averaged across trials and across category-neurons.

Next, we created population histograms for stimulus- and category-neurons at different learning stages, respectively. In the sequential paired-association task, the stimulus-neurons and the category-neurons produced different activity to stimuli. We found that when this model was in the early learning stage (at the 50th iteration) of the task, the neurons of both populations could show activity differences between the preferred and non-preferred stimuli during the first stimulus period and the first delay period. However, from the second stimulus period, these activity differences gradually disappeared for the both types of neurons (Figures 5B,C). When this model was in the final learning stage (at the 600th iteration) of the task, both stimulus-neurons and category-neurons show stronger activity to preferred stimuli than that to non-preferred stimuli in the whole trial (Figures 5D,E). The results indicated that although the information encoded by neurons would decay with time in the process of transmission, the neurons would gradually enhance the storage capacity of information and form working memory through learning.

In order to quantitatively measure activity-changes during the learning process, we calculated the category index for each stimulus- and category-neuron in each testing stage, respectively. The mean category index of the stimulus-neurons decreased gradually, and the mean category index of the category-neurons increased gradually during the learning process of the task (Figure 6A). It meant that category-neurons enhanced the ability to encode category information through learning; while stimulus-neurons did not exhibit the characteristic of enhanced ability to encode category information.

Second, we quantitatively characterize the ability of both types of neurons to encode stimulus information during the learning process. We computed the stimulus index for each neuron to denote response-differences to within-category stimuli (see Section Methods). The mean stimulus index of 58 stimulus-neurons increased gradually, and the mean stimulus index of 47 category-neurons kept relatively stable during the learning process of the task (Figure 6B). The result of the Mann-Whitney U test showed that there was a significant difference in the ability of two populations to discriminate within-category stimuli in the final learning stage ($p = 0.018$). For stimulus-neurons, although their ability for category coding decreased, their ability for stimulus coding obviously increased.

It was obvious that the ability of these two types of neurons to encode information was enhanced during the learning process of this model, and their activity also changed in different task periods. We further analyzed the characteristics of neurons encoding information in different task periods. Interestingly, the category-neurons show the strongest ability to encode category information in the first stimulus period, and the ability decreased after the first stimulus period. Even though, the mean category index of category-neurons was higher than that of the stimulus-neurons in each task period (Figure 6C). The stimulus-neurons showed the strongest ability to encode stimulus information in the first stimulus period, and this ability also decreased after the first stimulus period. But in each task period, the mean stimulus index of stimulus-neurons was higher than that of the category-neurons (Figure 6D).

Although the stimulus-neurons and category-neurons may play different roles in this model, we found that category-neurons encoded not only category information but also stimulus information (see Figure 5E, category-neurons could discriminate the three preferred stimuli). One question is whether the stimulus information found in the category-neurons is directly influenced by external stimuli? It was worth noting that in the policy network, sparse connections were used between neurons in the input and IILs. And only some neurons in the IIL directly received stimuli from the input layer (these neurons are denoted as directly connected neurons), while other neurons did not (those neurons that do not receive direct projections from the input layer as indirectly connected neurons). We analyzed the activity differences of the two groups of directly and indirectly connected neurons. In the first stimulus period, 54 (54/150; 36%) neurons in the IIL were directly connected with neurons in the input layer. Within them, 21 (21/54; 38.9%) neurons were identified as the category-neurons. And their mean category index was 0.715 (Figure 7A). In addition, 96 (96/150; 64%) neurons did not receive direct connections from the input layer. Among these 96 neurons, 26 (26/96; 27.1%) of them were identified as the category-neurons. And their mean category index was 0.745 (Figure 7B). The two groups of category-neurons had similar distributions of category indexes (see Figures 7A,B). Furthermore, we found that the two groups of neurons showed different learning curves of the category index (Figure 7C). The mean category index of directly connected neurons increased quickly in the early learning stage (at the 50th iteration) and changed slightly at later learning stages (from the 300th iteration to the 600th iteration). The mean category index of indirectly connected neurons increased obviously at different learning stages (from the 50th iteration to the 600th iteration). In the final learning stage (at the 600th iteration), the two groups of neurons showed similar category indexes (Mann-Whitney U test, $p = 0.250$).

We further calculated the stimulus indexes for the two groups of directly connected neurons, and indirectly connected neurons (Figure 7D). The mean stimulus index of directly connected neurons was significantly higher than that of indirectly connected neurons. The result of the Mann-Whitney U test showed that external stimuli directly affected the ability of category-neurons to discriminate between stimuli ($p = 0.002$). It indicated that the ability of neurons to encode category information during category learning was not directly influenced by external stimuli; whereas the ability of neurons to encode stimulus information was directly influenced by external stimuli.

Weight analysis of neurons in the network

It was found that the neurons in this model were capable of stimulus coding and category coding. This model updated weights during learning the sequential paired-association task. In general, the synaptic plasticity of neurons is crucial in constructing models (51, 52). This is because the information is exchanged between neurons with the help of synaptic connections, and the type of synapses (excitatory or inhibitory synapses) and their values affect the activity of neurons (53). At the computational level, excitatory synapses increase firing rates of neurons, while inhibitory synapses diminish their firing rates. So how does the interaction of excitatory and inhibitory synapses

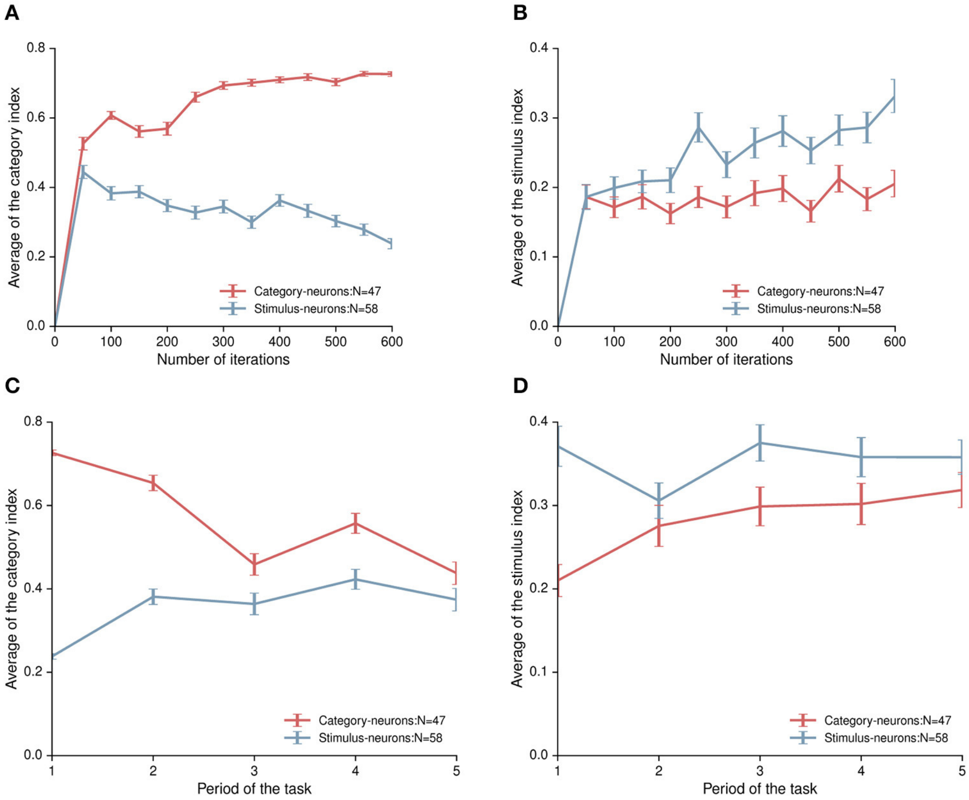


FIGURE 6
The category index and the stimulus index of category-neurons and stimulus-neurons. **(A)** The time course of the category index for category-neurons (the red curve) and stimulus-neurons (the blue curve) during the network model learning the task. **(B)** The time course of the stimulus index for category-neurons (the red curve) and stimulus-neurons (the blue curve) during the network learning the task. **(C)** The category indexes for category-neurons (the red curve) and stimulus-neurons (the blue curve) in five different task periods after the model learned the task. **(D)** The stimulus indexes for category-neurons (the red curve) and stimulus-neurons (the blue curve) in five different task periods. The number "1," "2," "3," "4," and "5" in the horizontal coordinates indicate the first stimulus period, the first delay period, the second stimulus period, the second delay period, and the third stimulus period, respectively.

affect the learning process of neurons? Therefore, we discussed the connection weights of neurons.

In the policy network of this model, the neurons in the input layer were sparsely connected to the neurons in the IIL with the probability of 0.2. Most neurons in the IIL could not directly receive the stimuli from the external environment. Here, among neurons in the IIL were sparsely connected with the probability of 0.1, and the neurons indirectly learned the stimuli from the external environment through information transmission. When this model was trained, we recorded the connection weights of neurons in the IIL, where positive values were excitatory weights and negative values were inhibitory weights. The connection weights of these neurons were Gaussian distribution (Figure 8A), and a balance mechanism was formed between excitatory weights and inhibitory weights.

Next, we asked a question whether the weight change between two connected neurons was correlated to the similarity of their activity-patterns. We would expect that neurons having more similar activity-patterns had stronger connection weights to form connection structures in the IIL during the learning process. To understand this problem, we selected every pair of connected neurons, and calculated the Pearson correlation coefficient of their activity-patterns in the first stimulus period. In addition, we also calculated the weight change (the difference between the weight at the end of training and the initial weight). Figure 8B shows scatter plots of the Pearson correlation coefficient and the weight change for all pairs of neurons. There is no correlation between them. Specifically, we used the same method to calculate the Pearson correlation coefficient and the weight change for the stimulus-neurons (Figure 8C) or for the category-neurons only (Figure 8D). Even within the same type of neurons, the similarity of their activity-patterns is not correlated with their weight changes. Although category-neurons were able to identify the category to which the stimuli belong, their activity-patterns were not directly influenced by the weights. As we know, the structure of recurrent neural networks is extremely complex. The neurons are not only involved in updating

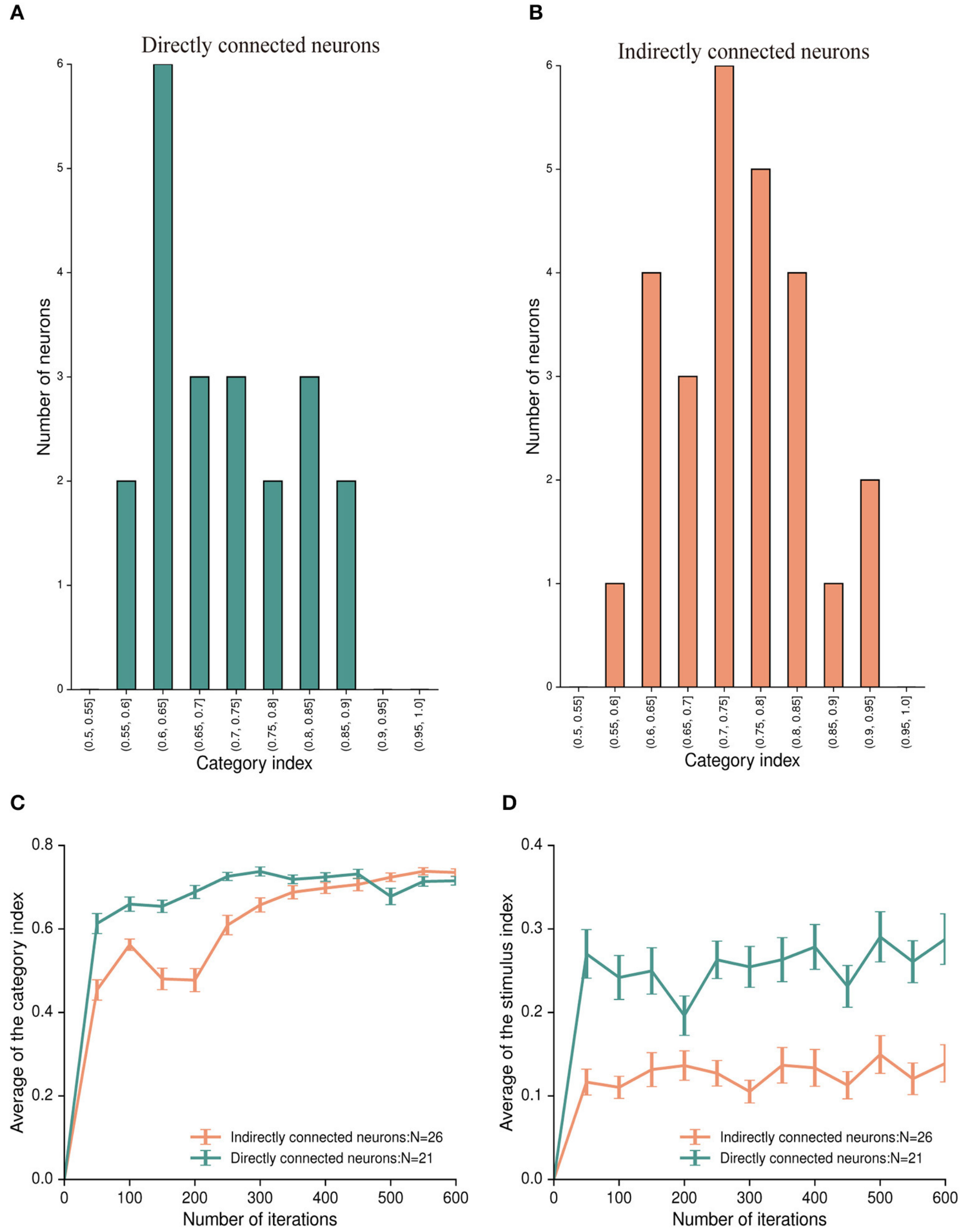


FIGURE 7
Category index and stimulus index for two types of category-neurons: directly connected neurons and indirectly connected neurons. **(A)** The distribution of category indexes of 21 category-neurons, which have direct connections from the input layer. The range of category index for these neurons is from 0.5 to 0.9. **(B)** The distribution of category indexes of 26 category-neurons, which have no direct connections from the input layer. The range of category index for these neurons is from 0.5 to 1. **(C)** The time course of category index for the directly connected neurons (the aquamarine curve) and the indirectly connected neurons (the salmon curve), respectively. **(D)** The time course of the stimulus index for the directly connected neurons (the aquamarine curve) and the indirectly connected neurons (the salmon curve), respectively.

weights during the learning process but also influenced by other factors, such as the decay of information and the importance of information, which meant that neurons produce similar activity performance as the result of the synergistic effect of multiple variables.

Neural activity related to action selection

Till now, we focused on analyzing neuronal activity in the first stimulus period, and found that majority of neurons encoded stimulus and category information. During the first

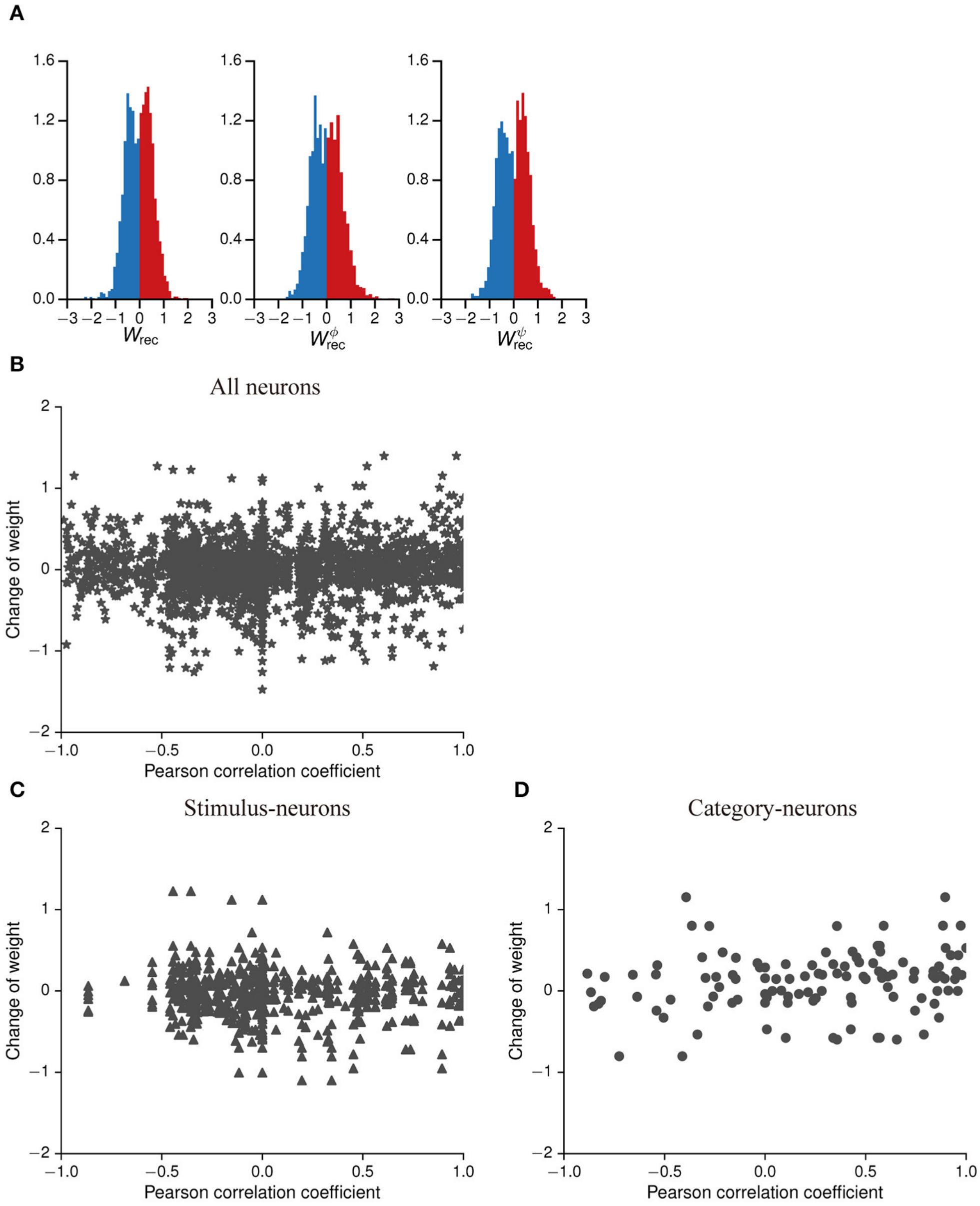


FIGURE 8
Distributions of weights in the policy network and the correlation analysis between activity-patterns and weight changes. **(A)** Frequency distribution histogram about the connection weights of neurons in the IIL of the policy network. The dark red bars denote excitatory weights, which are positive, and the dark blue bars denote inhibitory weights, which are negative. Left panel: weights of among neurons in the recurrent network; middle panel: weights of the update gates; right panel: weights of the reset gates. **(B–D)** show correlation analysis between the activity-pattern similarity of each pair of neurons (Pearson correlation coefficient) and their weight change. **(B)** All pairs of neurons that have connection in the IIL. **(C)** Pairs of connected neurons are selected only from stimulus-neurons. **(D)** Pairs of connected neurons are selected only from category-neurons.

stimulus period, the model had not to make a choice of action (left or right), there was no choice-related activity in this period. In the first decision period after the second stimuli offset, the model had to make a left or right choice. How was the choice-related information encoded in the IIL? In order to investigate this issue, we aligned neural activity at the first stimuli onset and sorted the activity into stimulus-position conditions (12 stimulus-position conditions, see Figure 9). We mainly found three types of activity-patterns in the first decision period (Figure 9). The first type of neurons showed no differential activity in response to the left position and right position in the first decision period, but showed differential activity to stimuli

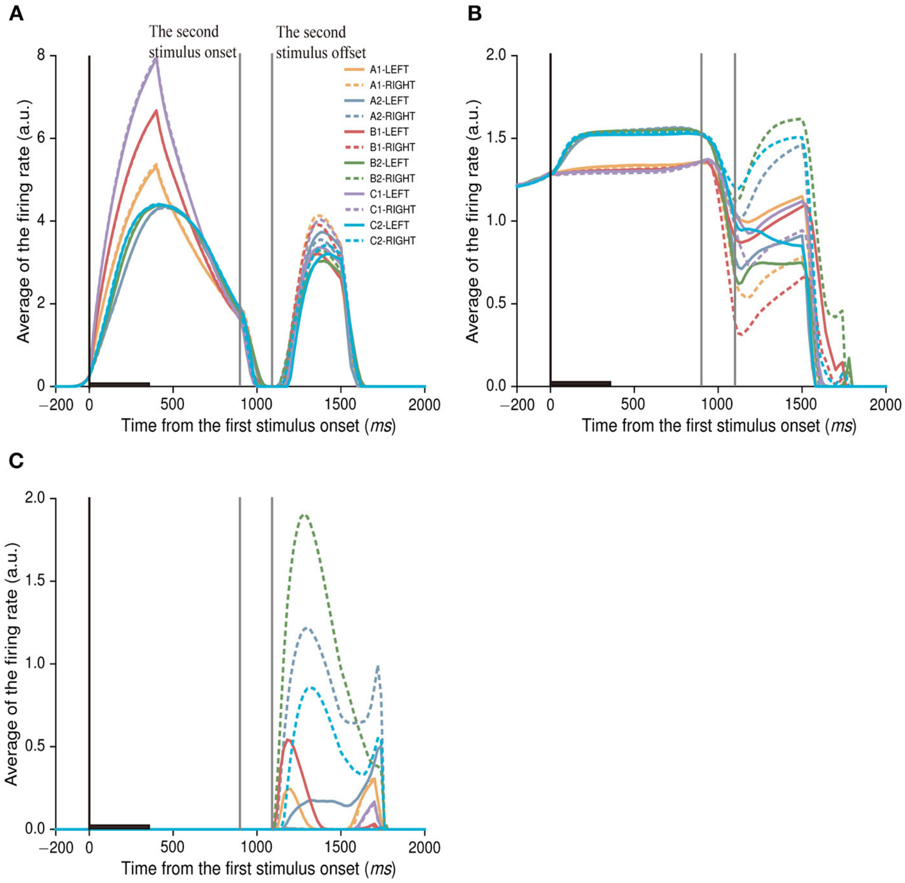


FIGURE 9
The activity of neurons related to action choices. Here, the two gray lines indicate the second stimulus period, after the second stimulus period, the network model chooses an action (left or right) during the first decision period. The activity of each neuron is aligned on the first stimulus onset, and is sorted with stimulus-position conditions. The same figure legends are used in **(A–C)**. **(A)** An example neuron shows only stimulus-related activity in the first stimulus and delay periods, no differential activity to actions (or positions). **(B)** An example of neuron shows not only category-related activity in the first stimulus and delay periods, but also activity related stimulus-action combinations after the second stimuli offset. **(C)** An example of neurons shows only stimulus-action related activity after the second stimuli offset, no stimulus-related activity either in the first stimulus period or in the delay period. The averaged firing rate of one neuron indicates its firing rates that are averaged across all trials.

in the first stimulus and delay periods (Figure 9A). This type of neurons may encode only stimulus-related information, no action-related activity. There were 21 neurons (21/150; 14%) that were classified into this type of neurons in the IIL. The second type of neurons could simultaneously encode stimulus-related information in the first stimulus and delay periods and stimulus-action combined information in the first decision period (Figure 9B). This type of neurons encoded information from pure stimulus-related information to stimulus-action information at different task periods. The number of this type of neurons was 75 (75/150; 50%) in the IIL. These neurons may contribute to transfer stimulus information into action information. The third type of neurons showed stimulus-action combined information only, no difference in response to the stimulus (Figure 9C). This neuron mainly discriminated between left-right actions. There were only 7 neurons found in the IIL. This type of neurons mainly contributed to action selection in the model. In addition, one third of neurons (47/150; 30.9%) showed no response during the whole trial (see Figure 4H; the firing rate of neurons was zero). The IIL neurons were able to encode stimulus information and position information, which were passed to neurons in the action output layer. The connection weights between neurons in the IIL and action output layer were dynamically adjusted during the training of this model. Finally, our model could learn the task.

Discussion

In this study, we demonstrated that the recurrent neural network using RL could learn the six stimulus-stimulus associative sequences. Through the trial-and-error method, the model first learned correct actions in the first choice and then in the second choice, a similar learning method was observed for the monkey in the same task. Various types of neural activity were found in the IIL in the first stimulus period. Some neurons encoded information of individual stimulus, and some neurons encoded category information of a group of stimuli that were associated together. These types of activity were also reported in the primate PFC in the sequential paired-association task (41). Actually, the stimulus-stimulus association task did not require the monkey and the model to form a categorical representation. However, some neurons in the PFC and in the IIL of the model did encode category information for associated stimuli. The categorical representation might help the monkey or the model accelerating the learning process. For example, the categorical representation of an associated stimuli enables them to easily select the target stimulus from a same category of the sample stimuli, without the requirement to memorize the specific target stimulus that is associated with the sample. Some neurons in the IIL also showed heterogenetic activity in different task periods (see Figure 9B). This type of heterogenetic activity-pattern was often observed in the PFC in different cognitive tasks (54). We found that almost half of model-neurons encoded stimulus (or category) information in the first stimulus and delay periods and encoded stimulus-position combined information in the decision periods. A few neurons encoded only stimulus-position combined information after the second stimuli offset. We did not find neurons that encoded pure position information (left or right action). The model learned stimulus information and then transferred it into stimulus-position combined information, and neurons in the action output layer integrated stimulus-position combined information to generate a correct action.

Many studies, including behavioral, neurophysiological, and fMRI experiments, suggest that the brain system learns categorical representations with a two-stage model of category learning (10, 20, 55). In the first stage, the sensory systems identify and represent stimulus information based on its physical properties (56). In the second stage, the associative brain areas encode meanings of a group of stimuli to form categorical representations. In our model, two types of neurons were found: stimulus-neurons and category-neurons. These neurons encode different aspects of task information, implying the model may learn category information with two different representations. Category-neurons encoded not only category information but also some stimulus information (see Figure 5E). Although those indirectly and directly connected neurons had the same level of category-indexes in the final learning stage, the former learned category information was slower than the latter (see Figure 7C). And the indirectly connected neurons also had significantly smaller stimulus-indexes than the directly connected neurons. These results indicate that inputs from the input layer may affect neurons in the IIL to learn category information. Further weight analysis suggested that stimulus-neurons or category-neurons did not form cluster or hierarchical structures in the IIL. The similarity of activity-patterns of a pair of neurons did not correlate to their weight changes (see Figure 8). In the current model, synaptic connections from the input layer to the IIL and within the IIL were sparse and random. A pre-determined connection structure in the IIL may help the model to learn representing stimulus information, category information and action information in a hierarchical manner.

The recurrent neural networks with the RL algorithm have been widely used to simulate behavior and neural activity of animals in cognitive tasks (57, 58). In this framework, our model is trained with the RL in a way similar to that the animals learn the cognitive task with trial and error. Model-neurons in the recurrent network appear complex and heterogenetic activity-patterns (see Figure 4). The RL algorithm plays a critical role in our model. Notably, the RL algorithm has a rich historical research background in machine learning (59–61). It has been reported that some brain areas, such as the PFC, the basal ganglia, and the dopamine system, implement RL to interact with the environment. Biologically, the PFC is critical in implementing strategies (62), and its neurons encode information about actions by adjusting strategies (63, 64). In addition, the PFC and the basal ganglia are interconnected to form a recurrent structure (65, 66). Specifically, the PFC and the striatum are closely linked (67). The dopamine is released from VTA (68) and SNc to the striatum and then acts on the PFC, and the information conveyed by dopamine is taken as the prediction error of the reward (69, 70), then the PFC adjusts the strategy based on the error signal from the striatum. Thus, the PFC is regarded as the policy network and the striatum is regarded as the value network (71).

It has been found that the brain areas, including the visual cortex, PFC, parietal cortex, premotor cortex, and basal ganglia are involved in the process of category learning (12, 21), in which the PFC neurons are more capable of encoding category information (72). Interestingly, in this model, we found that some model-neurons could encode category information for a group of stimuli that are associated each other in one chain, consistent with the finding that some PFC neurons encoded category information of those associated stimuli in the sequential paired-association task (10). While we don't know how exactly the PFC or the neuronal system to learn and form categorical representations for associated stimuli in the task, one possible way suggested by our current model is that the PFC and its related brain areas may implement deep RL to encode category information during learning the task.

It is well known that PFC neurons encode not only stimulus and category information but also reward information. For example, in the sequential paired-association task with an asymmetric reward schedule (47), PFC neurons showed strong responses to the stimulus that was associated with a large reward; when the reward amount was reversed for the same stimulus (large reward became small reward), these neurons responded slightly to the stimulus. This result demonstrated that the neural activity was affected by the amount of reward. We tried to make this model to learn the sequential paired-association task with an asymmetric reward schedule, but the activity-pattern of neurons was not influenced by the reward reversal. The possible reason is that we did not take into account the reward amount as a model parameter in this model. The reward amount is just considered as an error signal to modify connection weights in the policy network. Therefore, neurons in the IIL do not encode reward or stimulus-reward information. Remarkably, environmental stimuli as input information affect the neural activity in the model. If model-neurons receive different reward amounts as input information, their activity may reflect reward and stimulus-reward combined information, and this model might be able to complete the sequential paired-association task with an asymmetric reward schedule. This issue should be further investigated.

The simulated results in this study also demonstrate that the network model is able to encode categorical information efficiently. Hinaut and Dominey reported that some neurons in a three-layer recurrent neural network with randomly-initiated weights could encode the categorical structure of a set of behavioral sequences without the requirement to modify the weights (31). But the three-layer recurrent neural network did not encode category information efficiently, the percentage of such categorical neurons was very low (0.4% of total neurons) (31). Our model shows efficient ability to encode category information, almost one-third (47/150) of neurons have category-selectivity. However, there are still some limitations in the current model. For example, although it was found that neurons need the capacity of working memory to learn the sequential paired-association task, there was lack of detailed description of the working mechanism by which neurons store memory information. It is known that the brain has extremely complex neuronal circuits that are involved in category learning. Our model is a single-layer recurrent neural network, which has a relatively simple network structure. In the future, combination the long-short-term memory network (73, 74) with the asynchronous actor-critic algorithm (75, 76) should be included to construct models with multilayer recurrent structures to simulate functions of category-related neuronal circuits.

In summary, we use the framework of deep reinforcement learning (the recurrent network + reinforcement learning) to build the novel network model that is trained to learn the sequential paired-association task. This task requires the network model to make two sequential choices to learn stimulus-stimulus associations in one trial. Our new findings in this study are that the network model can perform the task correctly after being trained with the trial-and-error method, indicating that the model has the ability to learn the complex structure of the task, not just to learn simple stimulus-action or stimulus-reward associations as reported in previous studies (38, 42). More importantly, we found stimulus-neurons and category-neurons in the IIL of the policy network. These two types of neurons represent different aspects of task parameters, and their ability to encode category and stimulus information was strengthened during the learning process. The model neurons in the IIL show heterogenetic activity to encode information of the stimulus, category, action and their combinations. These responsive properties of neurons in the IIL are similar to activity-patterns observed in the primate PFC in the same task (41, 47), indicating the IIL could mimic functions of the PFC in the categorization tasks. The simulation results indicate that the recurrent neural network could learn the categorical representation for a group of stimuli in the matching-to-sample task (stimulus-stimulus associations) using the RL algorithm, without additional requirements such as the network structure, prior knowledge or specific categorical rules. Our results might provide a new way for understanding neuronal mechanisms underlying how the brain system learns category information.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

XP, YZ, and YW contributed to conception, design of the study, and wrote sections of the manuscript. YZ wrote the codes, performed simulating the model, and wrote the first draft of the manuscript. YZ and XP analyzed the data. All authors contributed to manuscript revision, read, and approved the submitted version.

Funding

This study was supported by the National Natural Science Foundation of China (Nos: 11972195, 12172132, 11802095) and Natural Science Foundation of Shanghai (No: 19zr1473100).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher’s note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Freedman DJ, Riesenhuber M, Poggio T, Miller EK. Visual categorization and the primate prefrontal cortex: neurophysiology and behavior. *J Neurophysiol.* (2002) 88:929–41. doi: 10.1152/jn.2002.88.2.929

2. Jee BD, Wiley J. Learning about the internal structure of categories through classification and feature inference. *Q J Exp Psychol.* (2014) 67:1786–807. doi: 10.1080/17470218.2013.871567

3. Ell SW, Smith DB, Deng R, Hélie S. Learning and generalization of within-category representations in a rule-based category structure. *Atten Percept Psychophys.* (2020) 82:2448–62. doi: 10.3758/s13414-020-02024-z

4. Ashby FG, Ennis JM, Spiering BJ. A neurobiological theory of automaticity in perceptual categorization. *Psychol Rev.* (2007) 114:632–56. doi: 10.1037/0033-295X.114.3.632

5. Tanaka S, Pan X, Oguchi M, Taylor JE, Sakagami M. Dissociable functions of reward inference in the lateral prefrontal cortex and the striatum. *Front Psychol.* (2015) 6:995. doi: 10.3389/fpsyg.2015.00995

6. Tsutsui K, Hosokawa T, Yamada M, Iijima T. Representation of functional category in the monkey prefrontal cortex and its rule-dependent use for behavioral selection. *J Neurosci.* (2016) 36:3038–48. doi: 10.1523/JNEUROSCI.2063-15.2016

7. Schlegelmilch R, Von Helversen B. The influence of reward magnitude on stimulus memory and stimulus generalization in categorization decisions. *J Exp Psychol Gen.* (2020) 149:1823–54. doi: 10.1037/xge0000747

8. Hosokawa T, Honda Y, Yamada M, Romero MDC, Iijima T, Tsutsui KI. Behavioral evidence for the use of functional categories during group reversal task performance in monkeys. *Sci Rep.* (2018) 8:15878. doi: 10.1038/s41598-018-33349-3

9. Zhou Y, Rosen MC, Swaminathan SK, Masse NY, Zhu O, Freedman DJ. Distributed functions of prefrontal and parietal cortices during sequential categorical decisions. *Elife.* (2021) 10:e58782. doi: 10.7554/eLife.58782

10. Pan X, Sakagami M. Category representation and generalization in the prefrontal cortex. *Eur J Neurosci.* (2012) 35:1083–91. doi: 10.1111/j.1460-9568.2011.07981.x

11. Jensen G, Kao T, Michaelcheck C, Borge SS, Ferrera VP, Terrace HS. Category learning in a transitive inference paradigm. *Mem Cognit.* (2021) 49:1020–35. doi: 10.3758/s13421-020-01136-z

12. Seger CA, Miller EK. Category learning in the brain. *Annu Rev Neurosci.* (2010) 33:203–19. doi: 10.1146/annurev.neuro.051508.135546

13. Nomura EM, Reber PJ. Combining computational modeling and neuroimaging to examine multiple category learning systems in the brain. *Brain Sci.* (2012) 2:176–202. doi: 10.3390/brainsci2020176

14. Meyers EM, Freedman DJ, Kreiman G, Miller EK, Poggio T. Dynamic population coding of category information in inferior temporal and prefrontal cortex. *J Neurophysiol.* (2008) 100:1407–19. doi: 10.1152/jn.90248.2008

15. Yamada Y, Kashimori Y. Neural mechanism of dynamic responses of neurons in inferior temporal cortex in face perception. *Cogn Neurodyn.* (2013) 7:23–38. doi: 10.1007/s11571-012-9212-2

16. Emadi N, Rajimehr R, Esteky H. High baseline activity in inferior temporal cortex improves neural and behavioral discriminability during visual categorization. *Front Syst Neurosci.* (2014) 8:218. doi: 10.3389/fnsys.2014.00218

17. Mansouri FA, Freedman DJ, Buckley MJ. Emergence of abstract rules in the primate brain. *Nat Rev Neurosci.* (2020) 21:595–610. doi: 10.1038/s41583-020-0364-5

18. Freedman DJ, Riesenhuber M, Poggio T, Miller EK. A comparison of primate prefrontal and inferior temporal cortices during visual categorization. *J Neurosci.* (2003) 23:5235–46. doi: 10.1523/JNEUROSCI.23-12-05235.2003

19. Davis T, Goldwater M, Giron J. From Concrete Examples to Abstract Relations: The Rostrolateral Prefrontal Cortex Integrates Novel Examples into Relational Categories. *Cereb Cortex.* (2017) 27:2652–70. doi: 10.1093/cercor/bhw099

20. Freedman DJ, Assad JA. Neuronal mechanisms of visual categorization: an abstract view on decision making. *Annu Rev Neurosci.* (2016) 39:129–47. doi: 10.1146/annurev-neuro-071714-033919

21. Viganò S, Borghesani V, Piazza M. Symbolic categorization of novel multisensory stimuli in the human brain. *Neuroimage.* (2021) 235:118016. doi: 10.1016/j.neuroimage.2021.118016

22. Seger CA. How do the basal ganglia contribute to categorization? Their roles in generalization, response selection, and learning via feedback. *Neurosci Biobehav Rev.* (2008) 32:265–78. doi: 10.1016/j.neubiorev.2007.07.010

23. Antzoulatos EG, Miller EK. Differences between neural activity in prefrontal cortex and striatum during learning of novel abstract categories. *Neuron.* (2011) 71:243–9. doi: 10.1016/j.neuron.2011.05.040

24. Valentin VV, Maddox WT, Ashby FG. Dopamine dependence in aggregate feedback learning: a computational cognitive neuroscience approach. *Brain Cogn.* (2016) 109:1–18. doi: 10.1016/j.bandc.2016.06.002

25. Ballard I, Miller EM, Piantadosi ST, Goodman ND, Mcclure SM. Beyond reward prediction errors: human striatum updates rule values during learning. *Cereb Cortex.* (2018) 28:3965–75. doi: 10.1093/cercor/bhx259

26. Soga M, Kashimori Y. Functional connections between visual areas in extracting object features critical for a visual categorization task. *Vision Res.* (2009) 49:337–47. doi: 10.1016/j.visres.2008.10.023

27. Chaisangmongkon W, Swaminathan SK, Freedman DJ, Wang XJ. Computing by robust transience: how the fronto-parietal network performs sequential, category-based decisions. *Neuron.* (2017) 93:1504–17.e4. doi: 10.1016/j.neuron.2017.03.002

28. Pinotsis DA, Siegel M, Miller EK. Sensory processing and categorization in cortical and deep neural networks. *Neuroimage.* (2019) 202:116118. doi: 10.1016/j.neuroimage.2019.116118

29. Cantwell G, Crossley MJ, Ashby FG. Multiple stages of learning in perceptual categorization: evidence and neurocomputational theory. *Psychon Bull Rev.* (2015) 22:1598–613. doi: 10.3758/s13423-015-0827-2

30. Bonnasse-Gahot L, Nadal JP. Categorical perception: a groundwork for deep learning. *Neural Comput.* (2022) 34:437–75. doi: 10.1162/neco_a_01454

31. Hinaut X, Dominey PF. A three-layered model of primate prefrontal cortex encodes identity and abstract categorical structure of behavioral sequences. *J Physiol Paris.* (2011) 105:16–24. doi: 10.1016/j.jphysparis.2011.07.010

32. Lee D, Seo H, Jung MW. Neural basis of reinforcement learning and decision making. *Annu Rev Neurosci.* (2012) 35:287–308. doi: 10.1146/annurev-neuro-062111-150512

33. Zhu H, Paschalidis IC, Hasselmo ME. Neural circuits for learning context-dependent associations of stimuli. *Neural Netw.* (2018) 107:48–60. doi: 10.1016/j.neunet.2018.07.018

34. Tsuda B, Tye KM, Siegelmann HT, Sejnowski TJ. A modeling framework for adaptive lifelong learning with transfer and savings through gating in the prefrontal cortex. *Proc Natl Acad Sci USA.* (2020) 117:29872–82. doi: 10.1073/pnas.2009591117

35. Schönberg T, Daw ND, Joel D, O'doherty JP. Reinforcement learning signals in the human striatum distinguish learners from nonlearners during reward-based decision making. *J Neurosci.* (2007) 27:12860–7. doi: 10.1523/JNEUROSCI.2496-07.2007

36. Mas-Herrero E, Sescousse G, Cools R, Marco-Pallarés J. The contribution of striatal pseudo-reward prediction errors to value-based decision-making. *Neuroimage.* (2019) 193:67–74. doi: 10.1016/j.neuroimage.2019.02.052

37. Lehnert L, Littman ML, Frank MJ. Reward-predictive representations generalize across tasks in reinforcement learning. *PLoS Comput Biol.* (2020) 16:e1008317. doi: 10.1371/journal.pcbi.1008317

38. Song HF, Yang GR, Wang XJ. Reward-based training of recurrent neural networks for cognitive and value-based tasks. *Elife.* (2017) 6:e21492. doi: 10.7554/eLife.21492

39. Zhang Z, Cheng Z, Lin Z, Nie C, Yang T. A neural network model for the orbitofrontal cortex and task space acquisition during reinforcement learning. *PLoS Comput Biol.* (2018) 14:e1005925. doi: 10.1371/journal.pcbi.1005925

40. Peters J, Schaal S. Reinforcement learning of motor skills with policy gradients. *Neural Netw.* (2008) 21:682–97. doi: 10.1016/j.neunet.2008.02.003

41. Pan X, Sawa K, Tsuda I, Tsukada M, Sakagami M. Reward prediction based on stimulus categorization in primate lateral prefrontal cortex. *Nat Neurosci.* (2008) 11:703–12. doi: 10.1038/nn.2128

42. Zhang X, Liu L, Long G, Jiang J, Liu S. Episodic memory governs choices: An RNN-based reinforcement learning model for decision-making task. *Neural Netw.* (2021) 134:1–10. doi: 10.1016/j.neunet.2020.11.003

43. Jordan ID, Sokół PA, Park IM. Gated recurrent units viewed through the lens of continuous time dynamical systems. *Front Comput Neurosci.* (2021) 15:678158. doi: 10.3389/fncom.2021.678158

44. Zhang Z, Cheng H, Yang T. A recurrent neural network framework for flexible and adaptive decision making based on sequence learning. *PLoS Comput Biol.* (2020) 16:e1008342. doi: 10.1371/journal.pcbi.1008342

45. Li Y, Yin B, Xi H. Partially observable Markov decision processes and performance sensitivity analysis. *IEEE Trans Syst Man Cybern B Cybern.* (2008) 38:1645–51. doi: 10.1109/TSMCB.2008.927711

46. Lillicrap TP, Santoro A. Backpropagation through time and the brain. *Curr Opin Neurobiol.* (2019) 55:82–9. doi: 10.1016/j.conb.2019.01.011

47. Pan X, Fan H, Sawa K, Tsuda I, Tsukada M, Sakagami M. Reward inference by primate prefrontal and striatal neurons. *J Neurosci.* (2014) 34:1380–96. doi: 10.1523/JNEUROSCI.2263-13.2014

48. Csete G, Bognár A, Csibri P, Kaposvári P, Sáry G. Aging alters visual processing of objects and shapes in inferotemporal cortex in monkeys. *Brain Res Bull.* (2015) 110:76–83. doi: 10.1016/j.brainresbull.2014.11.005

49. Cromer JA, Roy JE, Miller EK. Representation of multiple, independent categories in the primate prefrontal cortex. *Neuron.* (2010) 66:796–807. doi: 10.1016/j.neuron.2010.05.005

50. Roy JE, Riesenhuber M, Poggio T, Miller EK. Prefrontal cortex activity during flexible categorization. *J Neurosci.* (2010) 30:8519–28. doi: 10.1523/JNEUROSCI.4837-09.2010

51. Engel TA, Chaisangmongkon W, Freedman DJ, Wang XJ. Choice-correlated activity fluctuations underlie learning of neuronal category representation. *Nat Commun.* (2015) 6:6454. doi: 10.1038/ncomms7454

52. Ashby FG, Rosedahl L. A neural interpretation of exemplar theory. *Psychol Rev.* (2017) 124:472–82. doi: 10.1037/rev0000064

53. Di Maio V. The glutamatergic synapse: a complex machinery for information processing. *Cogn Neurodyn.* (2021) 15:757–81. doi: 10.1007/s11571-021-09679-w

54. Rigotti M, Barak O, Warden MR, Wang XJ, Daw ND, Miller EK, et al. The importance of mixed selectivity in complex cognitive tasks. *Nature.* (2013) 497:585–90. doi: 10.1038/nature12160

55. Goltstein PM, Reinert S, Bonhoeffer T, Hübener M. Mouse visual cortex areas represent perceptual and semantic features of learned visual categories. *Nat Neurosci.* (2021) 24:1441–51. doi: 10.1038/s41593-021-00914-5

56. Tasaka G, Ide Y, Tsukada M, Aihara T. Multimodal cortico-cortical associations induced by fear and sensory conditioning in the guinea pig. *Cogn Neurodyn.* (2022) 16:283–96. doi: 10.1007/s11571-021-09708-8

57. Han D, Doya K, Tani J. Self-organization of action hierarchy and compositionality by reinforcement learning with recurrent neural networks. *Neural Netw.* (2020) 129:149–62. doi: 10.1016/j.neunet.2020.06.002

58. Granato G, Cartoni E, Da Rold F, Mattera A, Baldassarre G. Integrating unsupervised and reinforcement learning in human categorical perception: a computational model. *PLoS ONE.* (2022) 17:e0267838. doi: 10.1371/journal.pone.0267838

59. Halici U. Reinforcement learning with internal expectation in the random neural networks for cascaded decisions. *Biosystems.* (2001) 63:21–34. doi: 10.1016/S0303-2647(01)00144-7

60. Chadderdon GL, Neymotin SA, Kerr CC, Lytton WW. Reinforcement learning of targeted movement in a spiking neuronal model of motor cortex. *PLoS ONE.* (2012) 7:e47251. doi: 10.1371/journal.pone.0047251

61. Lowet AS, Zheng Q, Matias S, Drugowitsch J, Uchida N. Distributional reinforcement learning in the brain. *Trends Neurosci.* (2020) 43:980–97. doi: 10.1016/j.tins.2020.09.004

62. Bussey TJ, Wise SP, Murray EA. The role of ventral and orbital prefrontal cortex in conditional visuomotor learning and strategy use in rhesus monkeys (*Macaca mulatta*). *Behav Neurosci.* (2001) 115:971–82. doi: 10.1037/0735-7044.115.5.971

63. Passingham RE, Toni I, Rushworth MF. Specialisation within the prefrontal cortex: the ventral prefrontal cortex and associative learning. *Exp Brain Res.* (2000) 133:103–13. doi: 10.1007/s002210000405

64. Yim MY, Cai X, Wang XJ. Transforming the choice outcome to an action plan in monkey lateral prefrontal cortex: a neural circuit model. *Neuron.* (2019) 103:520–32.e5. doi: 10.1016/j.neuron.2019.05.032

65. O'reilly RC, Frank MJ. Making working memory work: a computational model of learning in the prefrontal cortex and basal ganglia. *Neural Comput.* (2006) 18:283–328. doi: 10.1162/089976606775093909

66. Hélie S, Ell SW, Ashby FG. Learning robust cortico-cortical associations with the basal ganglia: an integrative review. *Cortex.* (2015) 64:123–35. doi: 10.1016/j.cortex.2014.10.011

67. Cantwell G, Riesenhuber M, Roeder JL, Ashby FG. Perceptual category learning and visual processing: An exercise in computational cognitive neuroscience. *Neural Netw.* (2017) 89:31–8. doi: 10.1016/j.neunet.2017.02.010

68. Chen M, Liu F, Wen L, Hu X. Nonlinear relationship between CAN current and Ca2+ influx underpins synergistic action of muscarinic and NMDA receptors on bursts induction in midbrain dopaminergic neurons. *Cogn Neurodyn.* (2022) 16:719–31. doi: 10.1007/s11571-021-09740-8

69. Daw ND, Gershman SJ, Seymour B, Dayan P, Dolan RJ. Model-based influences on humans' choices and striatal prediction errors. *Neuron.* (2011) 69:1204–15. doi: 10.1016/j.neuron.2011.02.027

70. Wang JX, Kurth-Nelson Z, Kumaran D, Tirumala D, Soyer H, Leibo JZ, et al. Prefrontal cortex as a meta-reinforcement learning system. *Nat Neurosci.* (2018) 21:860–8. doi: 10.1038/s41593-018-0147-8

71. Averbeck B, O'doherty JP. Reinforcement-learning in fronto-striatal circuits. *Neuropsychopharmacology.* (2022) 47:147–62. doi: 10.1038/s41386-021-01108-0

72. Mckee JL, Riesenhuber M, Miller EK, Freedman DJ. Task dependence of visual and category representations in prefrontal and inferior temporal cortices. *J Neurosci.* (2014) 34:16065–75. doi: 10.1523/JNEUROSCI.1660-14.2014

73. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput.* (1997) 9:1735–80. doi: 10.1162/neco.1997.9.8.1735

74. Alipour A, Beggs JM, Brown JW, James TW. A computational examination of the two-streams hypothesis: which pathway needs a longer memory? *Cogn Neurodyn.* (2022) 16:149–65. doi: 10.1007/s11571-021-09703-z

75. Wei Q, Wang L, Liu Y, Polycarpou MM. Optimal elevator group control via deep asynchronous actor-critic learning. *IEEE Trans Neural Netw Learn Syst.* (2020) 31:5245–56. doi: 10.1109/TNNLS.2020.2965208

76. Labao AB, Martija MAM, Naval PC. A3C-GS: adaptive moment gradient sharing with locks for asynchronous actor-critic agents. *IEEE Trans Neural Netw Learn Syst.* (2021) 32:1162–76. doi: 10.1109/TNNLS.2020.2980743

frontiers | Frontiers in Psychiatry

TYPE Original Research
PUBLISHED 24 November 2022
DOI 10.3389/fpsyt.2022.1022304

Check for updates

OPEN ACCESS

EDITED BY
Rubin Wang,
East China University of Science and Technology, China

REVIEWED BY
Jun Ma,
Lanzhou University of Technology, China
Jianwei Shen,
North China University of Water Resources and Electric Power, China

*CORRESPONDENCE
Mohd Effendi Ewan Mohd Matore
effendi@ukm.edu.my

SPECIALTY SECTION
This article was submitted to Computational Psychiatry, a section of the journal Frontiers in Psychiatry

RECEIVED 18 August 2022
ACCEPTED 27 October 2022
PUBLISHED 24 November 2022

CITATION
Sovey S, Osman K and Matore MEEM (2022) Gender differential item functioning analysis in measuring computational thinking disposition among secondary school students. *Front. Psychiatry* 13:1022304. doi: 10.3389/fpsyt.2022.1022304

Gender differential item functioning analysis in measuring computational thinking disposition among secondary school students

Saralah Sovey[1], Kamisah Osman[2] and Mohd Effendi Ewan Mohd Matore[3]*

[1]Sungai Ramal Secondary School, Bangi, Selangor, Malaysia, [2]Faculty of Education, Centre of STEM Enculturation, National University of Malaysia, Bangi, Selangor, Malaysia, [3]Faculty of Education, Research Centre of Education Leadership and Policy, National University of Malaysia, Bangi, Selangor, Malaysia

Computational thinking refers to the cognitive processes underpinning the application of computer science concepts and methodologies to the methodical approach and creation of a solution to a problem. The study aims to determine how students' cognitive, affective, and conative dispositions in using computational thinking are influenced by a gender. This study used a survey research design with quantitative approach. Five hundred thirty-five secondary school students were sampled using probability sampling with the Computational Thinking Disposition Instrument (CTDI). WINSTEPS version 3.71.0 software was subsequently employed to assess the Gender Differential item functioning (GDIF) including reliability and validity with descriptive statistics were employed to assess students' disposition toward practicing computational thinking. In addition to providing implications for the theory, the data give verifiable research that the CT disposition profile consists of three constructs. In addition, the demonstrated CTDI has good GDIF features, which may be employed to evaluate the efficacy of the application of CT in the Malaysian curriculum by measuring the level of CT in terms of the disposition profile of students.

KEYWORDS

computational thinking, disposition, Rasch model, gender differential item functioning, secondary school, student

Introduction

Computational thinking (CT) is a vital skill in any field. A number of researchers have proposed that CT serves as a stepping stone to more complex computational endeavors like programming (1). In particular, CT aid elementary school kids in conceptualizing computational reasoning. This is an ability that develops via repeated use (1). The educators viewed technology as a means to broaden their pupils' horizons and give them more agency in their own learning (2). Recent discussions have centered on the importance of introducing computer science to students in lower grades (3–6). The enthusiasm for CT in the classroom is understandable, but there are still many challenges that must be overcome. As a result, there is a growing expectation that teachers will be able to illustrate computational thinking by applying it to real-world scenarios that use computer technology.

Thus, an item is the fundamental unit of an instrument. The creation of items must be consistent and fair for all participants. DIF refers to a measurement instrument with multiple functions. It is being administered to a group of respondents with diverse demographic backgrounds but comparable abilities. Hambleton and Jones (7) suggest that a DIF-detected item's functions in various subgroups are dissimilar.

Consequently, the DIF analysis procedure identifies items that do not mirror similar functions when applied to a group of capabilities with parallel capabilities. Osterlind (8) states that item analysis entails observing items critically to reduce measurement error. Consequently, DIF analysis determines item validity (9). DIF endorsement in instrument construction is indicative of an instrument with high reliability. Siti Rahayah Ariffin (10), stated that DIF impacts the dependability of instruments. For composite measures to be unidimensional and the variable to be linear, the item scale values must be consistent across individuals and groups. Three DIF endorsement methods are Mantel-Haenszel (11, 12), Item Response Theory (13), and Rasch Models (14).

We are currently working on the next iteration of CT disposition instruments. Empirical evidence is essential for the creation of new statistical tools. As a result, the gender gap is one of the topics that has gained a lot of attention in academia, especially in the field of computer science education. Since many of the same ideas are used in both CT and computer science, a number of recent studies have looked into the disparity between the sexes in terms of CT proficiency.

From a neurological point of view, boys are a few weeks behind girls and remain behind girls until late adolescence (15). This developmental difference impacts their early school learning experiences and has impact throughout their education. Boys' fine motor skills develop slower than girls and they may have difficulty with handwriting tasks (16). Their language and fine motor skills fully mature about six years later than girls (17). However, the areas involved in targeting and spatial memory mature some four years earlier in boys than they do in girls (17). Although those differences are significant, it is important to examine how that information relates to developmental gender differences especially in CT. Recent studies in Cognitive Neurodynamics field also discussed several variables that cater interest such as decision-making (18), and brain activity patterns and mental (19). Other than that, the gender differences also reported in spirituality well-being (20) and mental fatigue (21). All these factors open the door to relate the Cognitive Neurodynamics with CT for developing better students in local context.

Thus, the very design of the brain and the resulting disparities in sensory perception and physical skills differ considerably between the sexes. Understanding those variances will assist instructors in providing a good and encouraging environment for their pupils, as well as promoting CT through teaching and learning.

Literature review

Computational thinking

In today's digital age, CT must be grasped quickly. CT is a kind of thinking that aligns with many 21st century abilities, such as problem-solving, creativity, and critical thinking (22). It derives from computer science and involves problem-solving, system design, and understanding people's behavior (5). CT refers to the cognitive process of problem solving (23) as a set of 21st century skills (24, 25) or the thought process involved in formulating problems and expressing solutions (26) and a set of problem-solving skills based on Computer Science (27). This encouraged researchers to perform more in-depth research on learning experiences and computational thinking methods (28). Researchers couldn't predict all difficulties before implementation (29).

CT is used from early childhood to university (30–32). The use of CT in formal education has taken numerous forms, including integration in computer science courses and embedding in math, science, and art (33, 34). (35) CT has also reached classrooms through robotics (36) and unplugged activities, such as board games or storybooks (37–39). While much has been said about demystifying CT pedagogy, research on evaluating CT skills and attitudes continues. To investigate systematic issues, it is indeed necessary to improve the attitude towards CT (40).

A review of the past five years (2016-2020) reveals that little research has been conducted on student CT disposition (41, 42). Correspondingly, attitudes affect CT as much as skills (5). CT's complexity inspires others to investigate further, implying a deeper understanding of CT as a disposition (43, 44). In the digital age, self-directed problem-solving instruction may no longer be adequate. This problem-solving method does not account for the willingness to incorporate these abilities. Thus, researchers propose that CT dispositions are crucial motivators for identifying complex real-world issues and developing effective solutions (45, 46). According to the National Research Council (NRC), specific thinking skills are associated with an innate desire to think and are constituents of specific thinking dispositions (47). Thus, good thinkers possess the ability to think and the disposition to think (48). It appears that a validated evaluation of CT dispositions is lacking.

Notwithstanding, there is a persistent desire in all fields to distinguish between various conceptualizations of CT measurement. Psychometric scales are one of the most often employed instruments for evaluating computational thinking (49–53). However, psychometric instruments are predicated on the notion that an individual provides accurate and comprehensive information (54). On the other hand, a western instrument might not be suitable for Malaysia, given the distinct cultural and geographic environments. As a result, an instrument that has already been verified might not be reliable in a different period, culture, or setting (55–57). Additionally, it can be difficult to compare data from various cultures and groups when studying attitudes (58).

Furthermore, it is a well-known fact that the issue of gender inequality in CT is coming up more frequently. Every student exhibits a different level of CT proficiency depending on their location, gender, and academic standing, as is well known (59). The CT of males and females is essentially the same. Although earlier studies (60–62) found no differences in CT skills between male and female students aged 15 to 18, gender inequalities still exist (6, 60). CT skillset is frequently correlated with mathematical reasoning, favoring male students (59). Researchers explained the contradictory results, suggesting that the task content might be to blame for the differences. For some tasks, boys or girls may find them more interesting (63). This implies that earlier research on gender issues has produced conflicting findings, demonstrating the need for additional study.

In Cognitive Neurodynamics aspect of human development, one of the important aspects for behavioral, cognitive, and neural sciences is related to decision-making (18). The complexity of real-life decision-making has the potential to be linked to one's person CT abilities. When a person is able to master CT well, then there is a possibility that their decision-making abilities also increase. CT may relate to brain activity patterns and mental. This point of discussion supports the findings of previous studies that there is a systematic link between brain activity patterns and spontaneously generated internal mental states (19).

In the context of this study, a person's gender in CT also encourages in effecting spirituality well-being. (20) in his study found the existence of a gender effect on spirituality and showed that alpha and theta brain signals increased in male students at the 30–35 age range; while this increase was slower at the 20-29 age range. External factors such as decision-making, brain activity patterns, internal mental, and spirituality also can be linked to a person's gender in CT differences. In addition, a study by Sadeghian et al. (21) also discusses mental fatigue based on gender. Their findings strengthen previous studies by showing the existence of a significant difference between the two groups of men and women for brain indicators with the alpha-1 index in men was higher than women and the average alpha-2 index in women was higher (both alpha indexes were to measure mental fatigue). This means that this difference in mental fatigue also has the potential to be linked to a person's CT ability according to gender.

However, most CT measurement methods focus on thinking skills rather than dispositions. The architecture of the CT disposition measurement model suggested in this paper is built on cognitive, affective, and conative. In this perspective, the study's importance can be viewed differently. It's important for developing a measurement tool's item pool and similar questions. It also helps create content for the most common components in modern literature. In many studies, limited tools such as perception-attitude scales, multiple choice tests, or just coding have been used to measure computational thinking (49, 51, 52, 64–66). This study established the Computational Thinking Disposition Instrument (CTDI) by considering several aspects of computational thinking.

Computational thinking disposition

The development of CT dispositions necessitates long-term involvement in computational techniques focused on the CT process (67). CT's psychological makeup remains a mystery to this day (52). When it comes to the internal impulse to act toward CT or respond in habitual but adaptive ways to people, events, or circumstances, the disposition is the term that describes it (68). While CT is most often regarded as a problem-solving process that emphasizes one's cognitive process and thinking skills (69, 70), more attention should be paid to the dispositions that students develop in CT education. CT dispositions refer to people's psychological status or attitudes when they are engaged in CT development (71). CT dispositions have recently been referred to as "confidence in dealing with complexity, a persistent working with difficulties, an ability to handle open-ended problems" (33, 72). Social psychologists describe dispositional traits as having an "attitudinal tendency" (73–75). Thoughtful dispositions, on the other hand, are often described in the context of critical thinking as a "mental frame or habit" (76). Furthermore, theorists argue that thinking is a collection of dispositions rather than knowledge or skill and that this is the case (77, 78).

Three psychological components comprise disposition: cognitive, affective, and conative. These three components of the mind are traditionally identified and studied by psychology (79–81). Information is encoded, perceived, stored, processed, and retrieved during cognition. A dispositional cognitive function is an individual's propensity to engage in cognitive mental activities such as perception, recognition, conception, judgment, and others. Affection is the emotional interpretation of sensations, data, or knowledge. People, things, and concepts are frequently associated with one's positive or negative relationships, and the question "How do I feel about this information or knowledge?" Self-actualization/self-satisfaction determines whether or not students feel successful after practicing CT in problem-solving exercises.

In contrast, conation refers to the relationship between knowledge, emotion, and behavior, which is ideally positive (rather than reactive or habitual) behavior (82, 83). Conative mental functions are "that aspect of mental activity that tends to develop into something else, such as the desire to act or a deliberate effort." Determination to an endeavor is a conative mental capacity. In this investigation, these attitudes and dispositions serve as theoretical entities. Different contexts and requirements necessitate distinct mental dispositions, according to the study's findings.

Due to the paucity of research and development on this topic, this study will make a substantial contribution to the body of knowledge as a result of its focus on computational thinking. Additionally, the tool gave an alternate perspective for evaluating students' success in the CT course. In response, we aim to take a psychometric approach to these challenges. On the other hand, the creation and development of our Computational Thinking Disposition Instrument is described, along with its descriptive statistics and dependability based on its administration to more than 500 Malaysian students. Consequently, the purpose of this work is to provide a novel instrument for assessing CT and to demonstrate the relationships between CT and other well-established psychological dimensions.

Research question

The purpose of this paper is to answer the following research question, which focuses on gender variations in attitudes concerning CT. Following the discussion on computational thinking disposition, a research question guides this paper:

1. To identify the existence of GDIF items in the Computational Thinking Disposition Instrument.

Methodology

Sample

The study employed a quantitative cross-sectional survey to collect and numerically analyze data to better comprehend the events under investigation (84). A self-administered online survey was used to collect the data, saving money, time, and effort. So, the data are almost ready for statistical analysis (85). The questionnaire survey was utilized since it is acceptable for a high sample size with a broad geographical coverage (86). This method also required respondents to check all boxes before submitting their responses, thereby minimizing data gaps. This study was conducted with the participation of 535 secondary school students with a background in computer science. Using probability sampling, samples were generated. Probability sampling employs a method of random selection that permits the estimation of sampling error, hence decreasing selection bias. Using a random sample, it is possible to describe quantitatively the relationship between the sample and the underlying population, giving the range of values, called confidence intervals, in which the true population parameter is likely to lie (87). Respondents were required to have a background in computer science, be willing to fill out questionnaires, and engage in online activities. The research was performed in October of 2020. Regarding ethical considerations, the student's permission to participate in this study was obtained prior to completing the questionnaire. Participation was voluntary and strictly anonymous. **Table 1** displays the demographic profile of the respondents.

TABLE 1 Demographic profile.

Demographic factor	Frequency	Percentage (%)	Total
Gender			
Male	247	46	535
Female	288	54	

Instrumentation

A Computational Thinking Disposition Instrument (CTDI) measures students' disposition in computational thinking. As was previously noted, three components were used to design the CTDI questionnaire. Sovey et al. (88) used factor analysis to demonstrate the items and validity constructions for the three constructs. EFA was the starting point of the investigation, then Rasch. The CTDI includes three demographic questions (gender, location, and prior knowledge) and 55 items in three dimensions that measure computational thinking disposition such as Cognitive (19 items), affective (17 items), and conative (19 items). All items had a 4-point Likert scale from strongly disagree (1) to strongly agree (4). Hence, there are recommendations that odd-numbered response scales should be avoided (89, 90). Dolnicar et al. (91) have explained that odd five-point Likert scales affect response styles that are biased, lack stability and take a long time to complete. The middle point scale category encourages a disproportionate number of responses (because the tendency to choose the middle scale is high). In the context of the study, the firmness of the respondent is considered an important basis in answering the items. Therefore, Sumintono and Widhiarso (92) have suggested not to provide a midpoint option. This argument is also supported by Wang et al. (93) who recommend that the midpoint scale not be used to obtain the views of Asian respondents. The scale is more appropriate compared to the conventional scoring method for the use of the Rasch model in this study. Ten pupils in total were then chosen for face validation. They were tasked with locating and cataloguing any unclear word or terminology. Additionally, they were permitted to share their thoughts on how to improve the questionnaires' quality in terms of font size and design so that the research sample could understand them more quickly. These 10 pupils were left out of the main study.

On the other hand, Rasch measurement model software WINSTEPS 3.73 was used to determine the instrument's validity and reliability. Rasch analysis (94) uses assumptions and a functional form to determine if a single latent trait drives questionnaire item responses. The Rasch model shows the assumed probability of participants' scale response patterns, which are added and tested against a probabilistic model (94, 95). The Rasch rating scale analysis model is used when a set of items share a fixed response rating scale format (e.g., Likert scale) and thresholds do not vary. Through its calibration of item difficulties and person abilities, the WINSTEPS software transformed raw ordinal data (Likert-type data), based on the frequency of response which appeared as probability, to logit (log odd unit) via the logarithm function, which assesses the overall fit of the instrument as well as person fit (96, 97). Rasch models are used in this study to determine gender. Bond and Fox (98, 99) propose three DIF indicators for groups that have been studied: (1) t value $\pm$ 2.0 ($-2.0 \leq t \leq 2.0$), (2) DIF Contrast $\pm$ 0.5 ($-0.5 \leq$ DIF Contrast ≤ 0.5), and (3) $p < 0.05$.

Person reliability and item reliability

According to Table 2, the "real" Person Reliability index (above 0.8) demonstrates that the consistency of individual responses was satisfactory (97). This indicates that the scale discriminates between individuals very well. This indicates that the likelihood of individuals responding to items was likely high. The same interpretation logic applies to Item Reliability measurements exceeding 0.90, which are also categorized as "very good" (100). High estimates of item reliability also indicate that the items define the latent variable very well (97). The CTDI may be considered a reliable instrument for various respondent groups.

Cronbach Alpha

The Cronbach Alpha coefficient value, as calculated by the Rasch Model, described the interaction between the 535 participants and the 55 items. According to Sumintono and Widhiarso's instrument quality criteria, a reliability score of more than 0.90 (Table 3) is considered "very good" (2014). This result indicates a high degree of interaction between the people and the items. An instrument is highly reliable if it has good psychometric internal consistency.

Person and item separation index

The person Separation index measures how well the CTDI can distinguish between 'Person abilities.' Item Separation index shows easy and difficult items' commonness (101). Wider is better. Bond and Fox (97) report that the item separation index is between 5.0 and 8.0, exceeding 2.0. Statistically, CTDI items could be divided into 5 to 8 endorsement levels. For respondents, a separation index above 2.0 is acceptable (102). Table 3 shows each construct's internal reliability. These criteria endorse the CTDI as a reliable instrument for assessing students' computational thinking disposition.

Results

Students' disposition towards computational thinking

Firstly, students' disposition towards computational thinking was analyzed. According to Table 4, among the three dimensions of disposition for computational thinking, students rated highest on affective, with a mean score of M = 2.76, SD = 2.08. However, lowest on the cognitive aspect, with a mean score of M = 2.48 and SD = 1.80. The results are summarized in Table 5.

Differences between students' demographic factors and computational thinking disposition

GDIF analysis is performed to determine biased items in the CTDI instrument. Table 3 shows the summary of GDIF items in each construct of CTDI. With the critical t-value set at 2.0 and the confidence level at 95%, nine items were identified as significant for GDIF, extending the analysis to identify the extreme level of GDIF that could exist in the items. Using the Rasch Model, we can predict which items are likely to exhibit biases and eliminate the most significant DIF-exhibiting items to improve test fairness. The negative t-value and GDIF size indicate that male students answered the questions more easily than female students. Four (44.4%) of the nine items indicating the existence of GDIF were easier for males, while five (55.6%) were easier for females. There is a sizeable proportion of items that appeal to both genders. The disparities between the sexes are minimal, and the business's direction is nonsystematic across all constructs. When the bias direction is not systematic, the moderately biased items are not problematic. The study revealed that item bias does not diminish the overall measurement accuracy and predictive validity of a test (103). As there is no benefit to removing these items, there should be a relatively high

TABLE 2 Reliability index and separation index.

	Respondent		Item		Cronbach Alpha
	Reliability index	Separation index	Reliability index	Separation index	
Cognitive	0.88	2.67	0.97	5.65	0.92
Affective	0.87	2.57	0.98	6.86	0.93
Conative	0.88	2.77	0.99	8.70	0.94

proportion of items in both ability groups that exhibit moderate DIF and have a low tendency to affect the instrument's quality.

Table 4 shows the results of GDIF analysis on cognitive items. Analysis revealed that only one of 19 items showed significant GDIF, K52. Item K52 ("I know the importance of citing reference sources for assignments undertaken") is easier to agree with by female students than male students. This item with a significant GDIF of 0.38 logits has a t-value of more than 2 ($t \geq 2.0$). Figure 1 shows the DIF plot using the DIF measure on the analysis of cognitive construct by gender where 1 indicate male students and 2 refers female students.

Table 6 shows the results of GDIF analysis on affective items. Analysis revealed that only one of 17 items showed significant GDIF, A1. Item A1 ("I do have a curiosity to explore new knowledge") is easier to agree with by female students than male students. This item with a significant GDIF of 0.42 logits has a t-value of more than 2 ($t \geq 2.0$). Figure 2 shows the DIF plot analysis of affective construct by gender (1 Male; 2 Female).

Table 7 shows the results of GDIF analysis on conative items. Analysis revealed that seven of 19 items showed significant GDIF, which are C3, C17, C22, C28, C31, C39, and C40. These items with significant GDIF ranging from 0.45 to 0.54 logits have a t-value of more than 2 ($t \geq 2.0$). Figure 3 shows the DIF plot DIF measure analysis of conative construct by gender. Item C3 ("I am willing to tolerate current group members during problem solving") is easier to agree with by female students than male students. Similarly, Item C17 ("I try to find the cause when a solution doesn't work") is easier to agree with by female students compared to male students. Additionally, Item C22 ("I diligently deal with a problem even beyond the allotted time") is easier to agree with by female students than male students. Conversely, Item C28 ("I am willing to take risks to solve a problem") is easier to agree with my male students than with female students. In addition, Item C31 ("I can adapt to the uncertainty of solving a problem") is easier to agree with male students than female students. Item C39 ("I have the courage to accept challenges to solve complex problems") is easier to agree with male students than female students. Item C40 ("I am confident that I understand the content of Computational Thinking") male students were more confident in understanding the content of computational thinking.

TABLE 3 Analysis of GDIF items.

Construct	Number of items	Items exhibit GDIF contrast	Direction of GDIF	
			Male	Female
Cognitive	19	1	-	1
Affective	17	1	-	1
Conative	19	7	4	3

Discussion

This advancement in science and technology has not had the same effect on men as it has on women. Differential item functioning (DIF) is present when two or more subgroups perform differently on a test item despite being matched on a measured construct. DIF analysis plays a crucial role in ensuring the equity and fairness of educational assessments since DIF-free instruments are regarded as equitable and fair for all participants. Consequently, the DIF study is a crucial procedure that aims to identify the item that does not demonstrate the same function when administered to students with the same ability but different backgrounds.

Nine out of the 55 items associated to gender in total do not fall within the acceptable range, hence it is suggested that they can be removed (99). The value was between 0.42 and 0.46, according to the DIF contrast results in Tables 4, 6, 7, while the t value was between 1.95 and 1.95 logits. The result is in agreement with the logit value of +0.5 to 0.5 determined for the DIF contrast for the Likert scale and the t value between 2 and +2. (97, 104). Apart from that, since the probability was higher than 0.05, these items did not include DIF (92). In general, geographic location, gender, and academic achievement affect a student's skillset (59). Gender differences are also a notable discussion in CT study (105). However, certain studies also indicated that males and females have similar CT. Despite no difference in CT between male and female 15-18-year-olds (61, 62), gender inequalities persist (6, 60).

Regarding the cognitive construct, K43 (I can change my mind to try something new while solving a problem) demonstrates that male students have superior cognitive abilities compared to female students. Boys are stronger at deductive and abstract reasoning, whereas girls are better at inductive and concrete (15). Boys reason from the general to the specific.

TABLE 4 GDIF analysis of cognitive construct.

Construct	Item	*t*-value	Probability (Welch)	GDIF size	The direction of item GDIF
Cognitive	K3	0.00	1.0000	0.00	Free
	K5	−0.89	0.3756	−0.15	Free
	K6	0.88	0.3795	0.15	Free
	K10	0.30	0.7679	0.05	Free
	K21	−0.14	0.8872	−0.02	Free
	K26	−1.11	0.2678	−0.20	Free
	K29	0.29	0.7717	0.05	Free
	K31	1.46	0.1460	0.26	Free
	K32	0.00	1.0000	0.00	Free
	K33	−0.47	0.6409	−0.09	Free
	K34	0.72	0.4748	0.12	Free
	K35	−1.35	0.1777	−0.23	Free
	K37	−1.23	0.2189	−0.21	Free
	K38	−0.96	0.3397	−0.16	Free
	K43	1.78	0.0760	0.30	Free
	K46	−0.79	0.4326	−0.13	Free
	K49	0.72	0.4724	0.12	Free
	K50	−1.29	0.1989	−0.22	Free
	K52	2.19	0.0287	0.38	Female

The colored cells mean that the items colored were biased to gender. The values are not fulfilled the *t*-value (±2.00) and the probability (Welch) (>0.05).

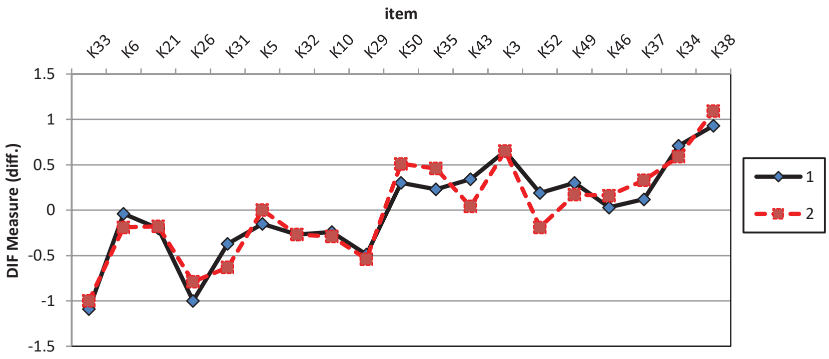


FIGURE 1
GDIF plot of cognitive items.

They employ concepts to solve problems. Male brains are 10-15% larger and heavier than female brains, according to study. Besides size, genders also differ in brain autonomy. Using brain mapping, researchers found that men have six times more gray matter connected to intelligence than women, but women have ten times more white matter. One study shows that gender-related differences in brain areas connect with IQ (106). This study and others show that males' inferior parietal lobes are larger. This lobe helps boys with spatial and mathematical reasoning. The left side of the brain, which controls language and verbal and written skills, develops sooner in girls, so they perform better in those areas (107). These results concur with Mouza et al. (108) 's conclusion that male students have a higher cognitive level of CT knowledge than female students. In addition, other studies have found that female students have limited computing knowledge and experience (109). According

TABLE 5 Descriptive statistics.

	Mean	Std. deviation
Cognitive	3.2410	0.4168
Affective	3.2903	0.4556
Conative	3.2771	0.4653

TABLE 6 GDIF analysis of affective construct.

Construct	Item	*t*-value	Probability (Welch)	GDIF size	The direction of item GDIF
Affective	A1	2.24	0.0256	0.42	Female
	A3	−1.95	0.0517	−0.33	Free
	A4	1.74	0.0833	0.32	Free
	A6	0.89	0.3751	0.17	Free
	A7	0.00	1.000	0.00	Free
	A9	−1.04	0.3001	−0.19	Free
	A10	−0.30	0.7617	−0.06	Free
	A11	0.31	0.7605	0.05	Free
	A14	0.25	0.8026	0.04	Free
	A18	0.13	0.8986	0.02	Free
	A19	0.00	1.000	0.00	Free
	A22	−1.37	0.1723	−0.23	Free
	A26	0.85	0.3964	0.15	Free
	A27	0.60	0.5475	0.11	Free
	A31	−0.09	0.6270	−0.28	Free
	A40	0.00	1.000	0.00	Free
	A41	0.27	0.7842	0.05	Free

The colored cells mean that the items colored were biased to gender. The values are not fulfilled the *t*-value (±2.00) and the probability (Welch) (>0.05).

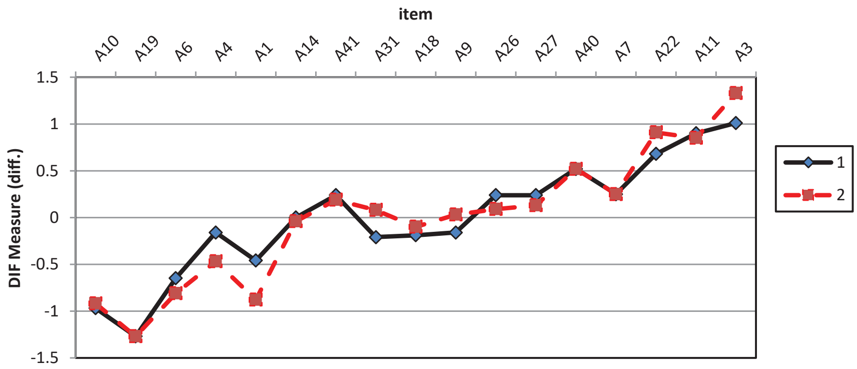


FIGURE 2
GDIF plot of affective items.

to research, males are typically more interested in information or knowledge than females (110). This may be due to the influence of culture and stereotypical socialization processes experienced by people beginning in childhood, as there are more men than women in these sectors (61). Lack of early experience and other obstacles contribute to girls' underrepresentation in this field (111).

Examining the affective construct, findings indicate that male students are more interested in practicing CT than female students for items A4 (I want to learn programming to apply Computational Thinking) and A1 (I have a curiosity to explore new knowledge). Similarly, Askar and Davenport (112) and Ozyurt and Ozyurt (113) found that male students have a greater sense of programming self-efficacy than female students. In addition, other studies have shown that the lack of female role models and differences in prior programming experience influence women's participation in computer science (114, 115). In addition, CT aptitude appears to be frequently linked to mathematical logic and favors male students (59). In addition, a study conducted outside of Malaysia revealed that male students were more familiar with technology and favored its use for learning (116). Female students typically require more time than male students to master CT (60). Atmatzidou and Demetriadis (60) reported that girls in the high school robotics STEM curriculum appeared to require more training time to attain the same skill level as boys in certain CT-specific aspects, such as decomposition.

TABLE 7 GDIF analysis of conative construct.

Construct	Item	*t*-value	Probability (Welch)	GDIF size	The direction of item GDIF
Conative	C1	1.95	0.0522	0.40	Free
	C3	2.32	0.0205	0.46	Female
	C6	0.52	0.6037	0.10	Free
	C7	1.56	0.1200	0.29	Free
	C8	1.70	0.0907	0.32	Free
	C10	0.00	1.0000	0.00	Free
	C17	2.52	0.0122	0.47	Female
	C19	0.89	0.3734	0.16	Free
	C21	0.87	0.3867	0.16	Free
	C22	2.49	0.0130	0.45	Female
	C28	−3.12	0.0019	−0.54	Male
	C29	0.54	0.5899	0.09	Free
	C31	−2.44	0.0150	−0.42	Male
	C38	−0.90	0.3680	−0.16	Free
	C39	−3.14	0.0018	−0.54	Male
	C40	−2.27	0.0236	−0.38	Male
	C42	−0.84	0.4020	−0.14	Free
	C44	−0.70	0.4870	−0.12	Free
	C45	−0.70	0.4836	−0.12	Free

The colored cells mean that the items colored were biased to gender. The values are not fulfilled the *t*-value (±2.00) and the probability (Welch) (>0.05).

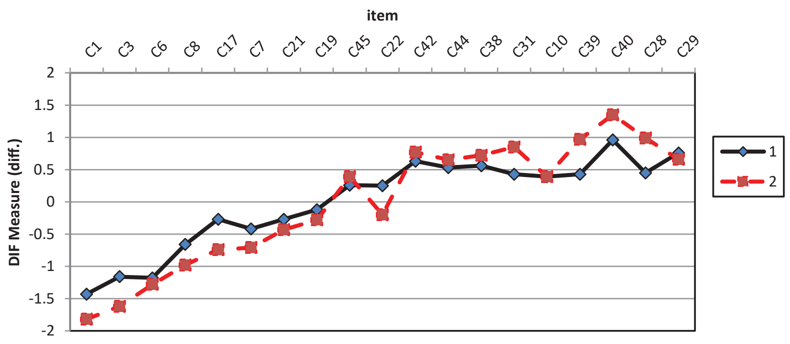


FIGURE 3
GDIF plot of conative items.

Moreover, in the conative construct, male students won item C17 (I try to find the cause when a solution does not work). According to neuroscience study (117–119), females' hippocampus develops faster and is larger than boys. Sequencing, vocabulary, reading, and writing are affected. Boys learn better through movement and visual experience because their cerebral cortex is more defined for spatial relationships. Girls favor collaborative activities where they exchange ideas with others, while guys prefer rapid, individualistic, kinesthetic, spatially-oriented, and manipulative-based activities (120). This fits with what Geary et al. (121) found: that male and female students have very different spatial and computational skills because male students are better at arithmetic reasoning. It is influenced by the fact that male students tend to be more intellectual, abstract, and objective. As a result, male students tend to understand issues through calculations, evaluate the compatibility between computational tools and techniques and challenges, and use computational strategies when solving problems. Besides, the additional information in Cognitive Neurodynamics context showed the existence of a gender effect on spirituality which reported that alpha and theta brain signals started increased in male according by age range (20). Computational thinking is breaking down complicated problems into steps that can be understood (called algorithms) and finding patterns that can be used to solve other problems (22). On the other hand, female students worked harder to find the first information. Still, female students tended to solve problems step by step, making it hard for them to find patterns or quick ways to solve problems. Women tend to be more careful, organized, and thorough than men (122, 123). Overall, the existing results proved that the existence of a significant difference between the two groups of men and women for brain indicators in the Cognitive Neurodynamics context.

Limitations and future directions

This study, like any other, has limitations. To begin with, this research is only focused on computational thinking disposition aspect. As a result, the CTDI instrument was built around three main elements in disposition. Thus, the first limitation is only Malaysian secondary school students were included in the study. Context can affect cultural differences. Researchers assert that context may explain the different results. Therefore, it is best to conduct larger-scale research with samples across Malaysia. This would increase the respondents' and research's demographic diversity.

DIF benchmarking could include in secondary schools. More research is needed to understand the differences in DIF item performance between groups, especially since computational thinking instruments are still being developed. We based our work on CT literature from various domains. The tool should apply to other fields. Replications in other countries would boost relevance in diverse nations. This study's construct validity came from a homogenous population. The scale must be validated with higher education, elementary, middle, and private school students. Comparing studies across tests can also improve psychometric assessment.

On the other hand, there are multiple areas for further research that stem from this study. Accordingly, future research should include different cultural groups to determine if the phenomenon is universal. East Asians and Westerners have different thinking patterns, according to Nisbett et al. (124). Westerners strongly prefer positivity, while Easterners have more varied preferences (125). This study will influence future analyses and improve item psychometrics. Further research can correlate personality traits. This instrument is DIF-analyzed in psychometrics to ensure it has not biased towards one measurement component as ethnic, socioeconomic status or age groups may contribute to the DIF.

Conclusion

This study assessed the effects of gender differences on disposition towards CT Using DIF analysis. Implementing a curriculum design to integrate STEM education with computational thinking to create an interdisciplinary approach presented a number of obstacles. A well-organized measuring instrument should be designed for long-term utilization. The information on gender clustering tendencies in answering GDIF items can help test developers create more fair achievement test items for students of different genders. The important practical implication is that the items selected from this study can be used as an alternative for self-evaluation and peer evaluation session for improvement purposes.

The findings of this study revealed a moderate level of CT disposition, which suggests the importance of making students aware of the evolution and rapid growth of CT discipline, and the availability of technological resources. The DIF analysis showed that there was a significant difference based on gender towards students' disposition for CT. Educators can use the data to identify students' strengths and weaknesses and plan more meaningful lessons. Girls and boys alike can flourish in their creative thinking if we teach them to focus on the process of CT and problem solving.

We need to acknowledge the fact that boys' and girls' perspectives on CT differ in significant ways. Differences in ability are not included in these categories. In order to encourage excellence in both sexes, educators must take into account the differences between males and females while planning lessons and activities. The necessity of devising engaging interventions and monitoring children's attentional and motivational elements during activities is illuminated by these findings, which have implications for educational practitioners and researchers. The study makes CT dispositions visible to the education community as path-opening invitations to explore CT and foster meaningful learning experiences.

Data availability statement

The original contributions presented in this study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Ethics statement

The studies involving human participants were reviewed and approved by Education Policy Planning and Research Division, Ministry of Education, Malaysia (Ethics approval number: KPM.600-3/2/3-eras (7906). Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

Author contributions

SS: conceptualization, methodology, formal analysis, investigation, writing—original draft preparation, and project administration. SS and MM: validation and data curation. SS, MM, and KO: resources and writing—review and editing.

MM: visualization and funding acquisition. MM and KO: supervision. All authors read and agreed to the published version of the manuscript.

Funding

This study was funded by the Faculty of Education, Universiti Kebangsaan Malaysia (UKM) with a Publication Reward Grant (GP-2021-K021854).

Acknowledgments

Special thanks and appreciation of gratitude to my supervisor, MM and KO, the Faculty of Education, UKM and Ministry of Education (MOE) for giving us the golden opportunity to conduct this wonderful project and guiding us from the beginning to the end, including each of the participants for their support during the data collection.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Denning PJ, Tedre M. *Computational Thinking*. Cambridge, MA: The MIT Press (2019). doi: 10.7551/mitpress/11740.001.0001

2. Rich KM, Yadav A, Larimore R. Teacher implementation profiles for integrating computational thinking into elementary mathematics and science instruction. *Educ Inf Technol.* (2020) 25:3161–88. doi: 10.1007/s10639-020-10115-5

3. Barr V, Stephenson C. Bringing computational thinking to K-12: what is involved and what is the role of the computer science education community. *ACM Inroads.* (2011) 2:48–54. doi: 10.1145/1929887.1929905

4. College Board. *Advanced Placement Computer Science Principles: Curriculum Framework*. (2014). Available online at: http://cms5.revize.com/revize/williamsvilleschools/Departments/Teaching%20&%20Learning/Mathematics/High%20School,%20Grades%209-12/ap-computer-science-principles-curriculum-framework.pdf (accessed December 1, 2021).

5. Wing JM. Computational thinking. *Commun ACM.* (2006) 49:33–5. doi: 10.1145/1118178.1118215

6. Yadav A, Mayfield C, Zhou N, Hambrusch S, Korb JT. Computational thinking in elementary and secondary teacher education. *ACM Trans Comput Educ.* (2014) 14:1–16. doi: 10.1145/2576872

7. Hambleton RK, Jones RW. Comparing classical test and item response theories and their applications to test development. *Educ Meas Issues Pract.* (1993) 12:38–47. doi: 10.1111/j.1745-3992.1993.tb00543.x

8. Osterlind SJ. *Constructing Test Items*. Boston, MA: Kluwer Academic Publishers (1989). doi: 10.1007/978-94-009-1071-3

9. Ackerman TA. A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *J Educ Meas.* (1992) 29:67–91. doi: 10.1111/j.1745-3984.1992.tb00368.x

10. Siti Rahayah Ariffin. *Inovasi dalam Pengukuran dan Penilaian Pendidikan [Innovation in Educational Measurement and Evaluation]*. Bangi: Fakulti Pendidikan UKM (2008).

11. Dodeen H. Stability of differential item functioning over a single population in survey data. *J Exp Educ.* (2004) 72:181–93. doi: 10.3200/JEXE.72.3.181-193

12. Stoneberg BD Jr. A Study of gender-based and ethnic based differential item functioning (DIF) in the Spring 2003 Idaho standards achievement tests applying the simultaneous bias test (SUBSET) and the mantel Haenszel chi-square test. *Intern Meas Stat.* (2004) 1–15. Available online at: https://files.eric.ed.gov/fulltext/ED483777.pdf

13. Maller SJ. Differential item functioning in the WISC-III: item parameters for boys and girls in the national standardization sample. *Educ Psychol Meas.* (2001) 61:793–817. doi: 10.1177/00131640121971527

14. Cauffman E, MacIntosh RA. Rasch differential item functioning analysis of the Massachusetts youth screening instrumentidentifying race and gender differential item functioning among juvenile offenders. *Educ Psychol Meas.* (2006) 66:502–321. doi: 10.1177/0013164405282460

15. Gurian M. *Boys and Girls Learn Differently: A Guide for Teachers and Parents I Michael Gurian and Patricia Henley with Terry Trueman.* 1st paperback ed. San Francisco, CA: Jossey-Bass (2002).

16. Pollack WS. *Real boys: Rescuing Our Sons From the Myths of Boyhood. William Pollack (1st Owl Books ed ed.).* New York, NY: Henry Holt & Company (1999).

17. Hanlon HW, Thatcher RW, Cline MJ. Gender differences in the development of EEG coherence in normal children. *Dev Neuropsychol.* (1999) 76:479. doi: 10.1207/S15326942DN1603_27

18. Vargas DV, Lauwereyns J.). Setting the space for deliberation in decision-making. *Cogn Neurodyn.* (2021) 15:743–55. doi: 10.1007/s11571-021-09681-2

19. Sen S, Daimi SN, Watanabe K, Takahashi K, Bhattacharya J, Saha G. Switch or stay? Automatic classification of internal mental states in bistable perception. *Cogn Neurodyn.* (2020) 14:95–113. doi: 10.1007/s11571-019-09548-7

20. MahdiNejad JED, Azemati H, Sadeghi habibabad A, Matracchi P. Investigating the effect of age and gender of users on improving spirituality by using EEG. *Cogn Neurodyn.* (2021) 15:637–47. doi: 10.1007/s11571-020-09654-x

21. Sadeghian M, Mohammadi Z, Mousavi SM. Investigation of electroencephalography variations of mental workload in the exposure of the psychoacoustic in both male and female groups. *Cogn Neurodyn.* (2022) 16:561–74. doi: 10.1007/s11571-021-09737-3

22. Yadav A, Hong H, Stephenson C. Computational thinking for all: pedagogical approaches to embedding 21st century problem solving in K-12 classrooms. *Tech Trends.* (2016) 60:565–8. doi: 10.1007/s11528-016-0087-7

23. García-Peñalvo FJ, Mendes AJ. *Exploring the Computational Thinking Effects in Pre-University Education*. Amsterdam: Elsevier (2018). doi: 10.1016/j.chb.2017.12.005

24. Mohaghegh M, McCauley M. Computational thinking: the skill set of the 21st century. *Int J Comput Sci Inf Technol.* (2016) 7:1524–30.

25. Curzon P, Black J, Meagher LR, McOwan PW. cs4fn.org: enthusing students about computer science. In: Hermann C, Lauer T, Ottmann T, Welte M editors. *Proceedings of Informatics Education Europe IV*. (2009). p. 73–80. Available online at: http://www.eecs.qmul.ac.uk/~pc/publications/2009/PCJBLRMPWCIEEIV2009preprint.pdf

26. Wing JM. *Computational Thinking Benefits Society. 40th Anniversary Blog of Social Issues in Computing*. New York, NY: Academic Press (2014).

27. Basso D, Fronza I, Colombi A, Pah C. Improving assessment of computational thinking through a comprehensive framework. In: *Proceedings of the 18th Koli Calling International Conference on Computing Education Research*. Koli (2018). doi: 10.1145/3279720.3279735

28. Eguchi A. Bringing robotics in classrooms. In: Khine M editor. *Robotics in STEM Education*. Cham: Springer (2017). p. 3–31. doi: 10.1007/978-3-319-57786-9_1

29. Belanger C, Christenson H, Lopac K. *Confidence and Common Challenges: The Effects of Teaching Computational Thinking to Students Ages 10-16*. Ph.D. Thesis. St Paul, MN: St.Catherine University Repository (2018).

30. Grover S, Pea R. Computational thinking in K–12: a review of the state of the field. *Educ Res.* (2013) 42:38–43. doi: 10.3102/0013189X12463051

31. Lyon JA, Magana JA. Computational thinking in higher education: a review of the literature. *Comput Appl Eng Educ.* (2020) 28:1174–89. doi: 10.1002/cae.22295

32. Fagerlund J, Häkkinen P, Vesisenaho M, Viiri J. Computational thinking in programming with scratch in primary schools: a systematic review. *Comput Appl Eng Educ.* (2021) 29:12–28. doi: 10.1002/cae.22255

33. Weintrop D, Beheshti E, Horn M, Orton K, Jona K, Trouille L, et al. Defining computational thinking for mathematics and science classrooms. *J Sci Educ Technol.* (2015) 25:1–21. doi: 10.1007/s10956-015-9581-5

34. Hickmott D, Prieto-Rodriguez E, Holmes K. A scoping review of studies on computational thinking in K–12 mathematics classrooms. *Digit Exp Math Educ.* (2018) 4:48–69. doi: 10.1007/s40751-017-0038-8

35. Bell J, Bell T. Integrating computational thinking with a music education context. *Inform Educ*. (2018) 17:151–66.

36. Ioannou A, Makridou E. Exploring the potentials of educational robotics in the development of computational thinking: a summary of current research and practical proposal for future work. *Educ Inf Technol.* (2018) 23:2531–44. doi: 10.1007/s10639-018-9729-z

37. Zhang L, Nouri J. A systematic review of learning computational thinking through scratch in K-9. *Comput Educ.* (2019) 141:103607. doi: 10.1016/j.compedu.2019.103607

38. Papadakis S. The impact of coding apps on young children computational thinking and coding skills. A literature reviews. *Front Educ.* (2021) 6:657895. doi: 10.3389/feduc.2021.657895

39. Stamatios P. Can preschoolers learn computational thinking and coding skills with Scratch Jr? A systematic literature review. *Int J Educ Reform*. (2022). doi: 10.1177/10567879221076077

40. Qiu RG. Computational thinking of service systems: dynamics and adaptiveness modeling. *Serv Sci.* (2009) 1:42–55. doi: 10.1287/serv.1.1.42

41. Haseski HI, Ilic U, Tugtekin U. Defining a new 21st-century skill-computational thinking: concepts and trends. *Int Educ Stud.* (2018) 11:29. doi: 10.5539/ies.v11n4p29

42. Jong MS, Geng J, Chai CS, Lin P. Development and predictive validity of the computational thinking disposition questionnaire. *Sustainability.* (2020) 12:4459. doi: 10.3390/su12114459

43. Tang K, Chou T, Tsai C. A content analysis of computational thinking research: an international publication trends and research typology. *Asia Pac Educ Res.* (2019) 29:9–19. doi: 10.1007/s40299-019-00442-8

44. Wing JM. Computational thinking and thinking about computing. *Philos Trans R Soc A Math Phys Eng Sci.* (2008) 366:3717–25. doi: 10.1098/rsta.2008.0118

45. Abdullah N, Zakaria E, Halim L. The effect of a thinking strategy approach through visual representation on achievement and conceptual understanding in solving mathematical word problems. *Asian Soc Sci.* (2012) 8:30. doi: 10.5539/ass.v8n16p30

46. Denning PJ. The profession of IT beyond computational thinking. *Commun ACM.* (2019) 52:28–30. doi: 10.1145/1516046.1516054

47. NRC. *Report of A Workshop on The Pedagogical Aspects of Computational Thinking*. Washington, DC: National Academies Press (2011).

48. Barr D, Harrison J, Conery L. Computational thinking: a digital age skill for everyone. *Learn Lead Technol.* (2011) 38:20–3.

49. Haseski HI, Ilic U. An investigation of the data collection instruments developed to measure computational thinking. *Inform Educ.* (2019) 18:297. doi: 10.15388/infedu.2019.14

50. Poulakis E, Politis P. Computational thinking assessment: literature review. In: Tsiatsos T, Demetriadis S, Mikropoulos A, Dagdilelis V editors. *Research on E-Learning and ICT in Education*. Berlin: Springer (2021). p. 111–28. doi: 10.1007/978-3-030-64363-8_7

51. Roman-Gonzalez M, Moreno-Leon J, Robles G. Combining assessment tools for a comprehensive evaluation of computational thinking interventions. In: Kong SC, Abelson H editors. *Computational Thinking Education*. Singapore: Springer (2019). p. 79–98. doi: 10.1007/978-981-13-6528-7_6

52. Tang X, Yin Y, Lin Q, Hadad R, Zhai X. Assessing computational thinking: a systematic review of empirical studies. *Comput Educ.* (2020) 148:103798. doi: 10.1016/j.compedu.2019.103798

53. Varghese VV, Renumol VG. Assessment methods and interventions to develop computational thinking—A literature review. In: *Proceedings of the 2021 International Conference on Innovative Trends in Information Technology (ICITIIT)*. Piscataway, NJ: IEEE (2021). p. 1–7.

54. Alan Ü. *Likert Tipi Olçeklerin ¨ Çocuklarla kullanımında yanıt kategori sayısının psikometrik Ozelliklere ¨ etkisi [effect of Number of Response Options on Psychometric Properties of Likert-Type Scale for Used With Children]*. Ph.D Thesis. Ankara: Hacettepe University (2019).

55. Beaton DE, Bombardier C, Guillemin F, Ferraz MB. Guidelines for the process of cross-cultural adaptation of self-report measures. *Spine.* (2000) 25:3186–91. doi: 10.1097/00007632-200012150-00014

56. Boynton PM, Greenhalgh T. Selecting, designing, and developing your questionnaire. *BMJ.* (2004) 328:1312–5. doi: 10.1136/bmj.328.7451.1312

57. Reichenheim ME, Moraes CL. Operacionalização de adaptação transcultural de instrumentos de aferição usados em epidemiologia [Operationalizing the cross-cultural adaptation of epidemiological measurement instruments]. *Rev Saude Publ.* (2007) 41:665–73. doi: 10.1590/S0034-89102006005000035

58. Gjersing L, Caplehorn JR, Clausen T. Cross-cultural adaptation of research instruments: language, setting, time, and statistical considerations. *BMC Med Res Methodol.* (2010) 10:13. doi: 10.1186/1471-2288-10-13

59. Chongo S, Osman K, Nayan NA. Level of computational thinking skills among secondary science students: variation across gender and mathematics achievement skills among secondary science students. *Sci Educ Int.* (2020) 31:159–63. doi: 10.33828/sei.v31.i2.4

60. Atmatzidou S, Demetriadis S. Advancing students' computational thinking skills through educational robotics: a study on age and gender relevant differences. *Robot Autonom Syst.* (2016) 75:661–70. doi: 10.1016/j.robot.2015.10.008

61. Espino EE, González CS. Influence of gender on computational thinking. In: *Proceedings of the XVI International Conference on Human-Computer Interaction*. Vilanova i la Geltru (2015). p. 119–28. doi: 10.1145/2829875.2829904

62. Oluk A, Korkmaz Ö. Comparing students' scratch skills with their computational thinking skills in different variables. *Int J Modern Educ Comput Sci.* (2016) 8:1–7. doi: 10.5815/ijmecs.2016.11.01

63. Izu C, Mirolo C, Settle A, Mannila L, Stupurienë G. Exploring bebras tasks content and performance: a multinational study. *Inf Educ.* (2017) 16:39–59. doi: 10.15388/infedu.2017.03

64. de Araujo ALSO, Andrade WL, Guerrero DDS. *A Systematic Mapping Study on Assessing Computational Thinking Abilities*. Piscataway, NJ: IEEE (2016). p. 1–9. doi: 10.1109/FIE.2016.7757678

65. Kong SC. Components and methods of evaluating computational thinking for fostering creative problem-solvers in senior primary school education. In: Kong SC, Abelson H editors. *Computational Thinking Education*. Singapore: Springer (2019). p. 119–41. doi: 10.1007/978-981-13-6528-7_8

66. Martins-Pacheco LH, von Wangenheim CAG, Alves N. Assessment of computational thinking in K-12 context: educational practices, limits and possibilities-a systematic mapping study. *Proceedings of the 11th international conference on computer supported education (CSEDU 2019)*. Heraklion: (2019). p. 292–303. doi: 10.5220/0007738102920303

67. Brennan K, Resnick M. New frameworks for studying and assessing the development of computational thinking. In: *Paper Presented at the Annual Meeting of the American Educational Research Association (AERA)*. Vancouver, BC (2012).

68. CSTA. *CSTA K–12 Computer Science Standards (revised 2017)*. (2017). Available online at: http://www.csteachers.org/standards (accessed December 13, 2021).

69. Lee I, Martin F, Denner J, Coulter B, Allan W, Erickson J, et al. Computational thinking for youth in practice. *ACM Inroads.* (2011) 2:32–7. doi: 10.1145/1929887.1929902

70. Brennan K, Resnick M. New frameworks for studying and assessing the development of computational thinking. In: *Proceedings of the 2012 Annual Meeting of the American Educational Research Association*. Vancouver, BC (2012). p. 25.

71. Halpern DF. Teaching critical thinking for transfer across domains: disposition, skills, structure training, and metacognitive monitoring. *Am Psychol.* (1998) 53:449. doi: 10.1037/0003-066X.53.4.449

72. Woollard J. CT driving computing curriculum in England. *CSTA Voice.* (2016) 12:4–5.

73. Facione PA. The disposition toward critical thinking: its character, measurement, and relationship to critical thinking skill. *Inform Logic.* (2000) 20:61–84. doi: 10.22329/il.v20i1.2254

74. Facione PA, Facione NC, Sanchez CA. *Test Manual for the CCTDI*. 2nd ed. Berkeley, CA: The California Academic Press (1995).

75. Sands P, Yadav A, Good J. Computational thinking in K-12: in-service teacher perceptions of computational thinking. In: Khine M editor. *Computational Thinking in the STEM Disciplines*. Cham: Springer (2018). doi: 10.1007/978-3-319-93566-9_8

76. Beyer BK. *Critical Thinking*. Bloomington, IN: Phi Delta Kappa Educational Foundations (1995).

77. Beyer BK. *Developing A Thinking Skills Program*. Boston, MA: Allyn and Bacon Inc (1988).

78. Norris SP, Ennis RH. *Evaluating Critical Thinking*. Tulsa, OK: Midwest Publications (1989).

79. Hilgard ER. The trilogy of mind: cognition, affection, and conation. *J Hist Behav Sci.* (1980) 16:107–17. doi: 10.1002/1520-6696(198004)16:2<107::AID-JHBS2300160202>3.0.CO;2-Y

80. Huitt W, Cain S. *An Overview of the Conative Domain. In Educational Psychology Interactive*. Valdosta, GA: Valdosta State University (2005). p. 1–20.

81. Tallon A. *Head and Heart: Affection, Cognition, Volition as Triune Consciousness*. The Bronx, NY: Fordham University (1997).

82. Baumeister REF, Bratslavsky E, Muraven M, Tice DM. Ego depletion: is the active self a limited resource. *J Personal Soc Psychol.* (1998) 74:1252–65. doi: 10.1037/0022-3514.74.5.1252

83. Emmons RA. Personal strivings: an approach to personality and subjective well-being. *J Pers Soc Psychol.* (1986) 51:1058–68. doi: 10.1037/0022-3514.51.5.1058

84. Gay LR, Mills GE. *Educational Research: Competencies for Analysis and Applications*. 12th ed. Hoboken, NJ: Merrill Prentice-Hall (2018).

85. Hair JF, Celsi MW, Harrison DE. *Essentials of Marketing Research*. 5th ed. New York, NY: McGraw-Hill Education (2020).

86. Creswell JW, Creswell JD. *Research Design: Qualitative, Quantitative, and Mixed Methods approach*. 5th ed. Los Angeles, CA: SAGE (2018).

87. Tyrer S, Heyman B. Sampling in epidemiological research: issues, hazards and pitfalls. *BJPsych Bull.* (2016) 40:57–60. doi: 10.1192/pb.bp.114.050203

88. Sovey S, Osman K, Matore MEE. Exploratory and confirmatory factor analysis for disposition levels of computational thinking instrument among secondary school students. *Eur J Educ Res.* (2022) 11:639–52. doi: 10.12973/eu-jer.11.2.639

89. Fisher JWP. Survey design recommendations. *Rasch Meas Trans.* (2006) 20:1072–4.

90. Yount R. *Research Design & Statistical Analysis in Christian Ministry*. 4th ed. Fort Worth, TX: Department of Foundations of Education (2006).

91. Dolnicar S, Grun B, Leisch F, Rossiter J. Three good reasons NOT to use five- and seven-point likert items. In: *Proceedings of the 21st CAUTHE National Conference*. Adelaide, SA: University of Wollongong (2011). p. 8–11.

92. Sumintono B, Widhiarso W. *Aplikasi Model Rasch untuk Penelitian Ilmu-Ilmu Sosial (Edisi revisi) [Application of Rasch Modelling in Social Science Research, Revised Edition]*. Cimahi: Trimkom Publishing House (2014).

93. Wang R, Hempton B, Dugan JP, Komives SR. Cultural differences: why do asians avoid extreme responses? *Surv Pract.* (2008) 1:1–7. doi: 10.29115/SP-2008-0011

94. Rasch G. *Probabilistic Models for Some Intelligence and Attainment Tests*. San Diego, CA: MESA Press (1960).

95. Boone WJ. Rasch analysis for instrument development: why, when, and how? *CBE Life Sci Educ.* (2016) 15:rm4. doi: 10.1187/cbe.16-04-0148

96. Linacre JM. *A User's Guide to WINSTEPS: Rasch Model Computer Programs*. San Diego, CA: MESA Press (2012).

97. Bond TG, Fox CM. *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*. 3rd ed. New York, NY: Routledge (2015).

98. Bond TG, Fox CM. *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*. New Jersey, NJ: Lawrence Erlbaum Associates Publishers (2001).

99. Bond TG, Fox CM. *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*. 2nd ed. New Jersey, NJ: Routledge (2007).

100. Fisher JWP. Rating scale instrument quality criteria. *Rasch Meas Trans.* (2007) 21:1095.

101. Boone WJ, Staver JR, Yale MS. *Rasch Analysis in the Human Sciences*. Dordrecht: Springer (2014). doi: 10.1007/978-94-007-6857-4

102. Fox CM, Jones JA. Uses of Rasch modeling in counselling psychology research. *J Couns Psychol.* (1998) 45:30–45. doi: 10.1037/0022-0167.45.1.30

103. Sheppard R, Han K, Colarelli S, Dai G, King D. Differential item functioning by sex and race in the hogan personality inventory. *Assessment.* (2006) 13:442–53. doi: 10.1177/1073191106289031

104. Lai JS, Eton DT. Clinically meaningful gaps. *Rasch Meas Trans.* (2002) 15:850.

105. Gunbatar MS, Karalar H. Gender differences in middle school students' attitudes and self-efficacy perseptions towards mblock programming. *Eur J Educ Res.* (2018) 7:925–33. doi: 10.12973/eu-jer.7.4.925

106. Kaufmann C, Elbel G. Frequency dependence and gender effects in visual cortical regions involved in temporal frequency dependent pattern processing. *Hum Brain Mapp.* (2001) 14:28–38. doi: 10.1002/hbm.1039

107. Gabriel P, Schmitz S. *Gender differences in occupational distributions among workers. Monthly labor review (June)*. Washington, DC: U.S. Bureau of Labor Statistics Division of Information and Marketing Services (2007).

108. Mouza C, Marzocchi A, Pan Y, Pollock L. Development, implementation, and outcomes of an equitable computer science after-school program: findings from middle-school students. *Res Technol Educ.* (2016) 48:84–104. doi: 10.1080/15391523.2016.1146561

109. Hur JW, Andrzejewski CE, Marghitu D. Girls and computer science: Experiences, perceptions, and career aspirations. *Comput Sci Educ.* (2017) 27:100–20. doi: 10.1080/08993408.2017.1376385

110. Paderewski P, García M, Gil R, González C, Ortigosa EM, Padilla-Zea N. Acercando las mujeres a la ingeniería: iniciativas y estrategias que favorecen su inclusión. *En XVI Congreso Internacional de Interacción Persona-Ordenador. Workshop Engendering Technology (II)*. Vilanova i la Geltrú: Asociación Interacción Persona Ordenador (AIPO) (2015). p. 319–26.

111. Cheryan S, Ziegler SA, Montoya AK, Jiang L. Why are some STEM fields more gender-balanced than others? *Psychol Bull.* (2017) 145:1–35. doi: 10.1037/bul0000052

112. Askar P, Davenport D. An investigation of factors related to self-efficacy for java programming among engineering students. *Turk Online J Educ Technol.* (2009) 8:26–32.

113. Ozyurt O, Ozyurt H. A study for determining computer programming students' attitudes towards programming and their programming self-efficacy. *J Theor Pract Educ.* (2015) 11:51–67. doi: 10.17718/tojde.58767

114. Beyer S. Why are women underrepresented in computer science? Gender differences in stereotypes, self-efficacy, values, and interests and predictors of future CS course-taking and grades. *Comput Sci Educ.* (2014) 24:153–92. doi: 10.1080/08993408.2014.963363

115. Wilson BC. A study of factors promoting success in computer science, including gender differences. *Comput Sci Educ.* (2002) 12:141–64. doi: 10.1076/csed.12.1.141.8211

116. Naresh B, Reddy BS, Pricilda U. A study on the relationship between demographic factors and e-learning readiness among students in higher education. *Sona Glob Manag Rev.* (2016) 10:1–11.

117. Sousa DA, Tomlinson CA. *Differentiation and the Brain: How Neuroscience Supports the Learner-Friendly Classroom*. Bloomington, IN: Solution Tree Press (2011).

118. King K, Gurian M. The brain–his and hers. *Educ Leadersh.* (2006) 64:59.

119. Bonomo V. Gender matters in elementary education research-based strategies to meet the distinctive learning needs of boys and girls. *Educ Horiz.* (2001) 88:257–64.

120. Gurian M. *The Boys and Girls Learn Differently Action Guide for Teachers*. San Francisco, CA: Jossey-Bass (2003).

121. Geary DC, Saults SJ, Liu F, Hoard MK. Sex differences in spatial cognition, computational fluency, and arithmetical reasoning. *J Exp Child Psychol.* (2000) 77:337–53. doi: 10.1006/jecp.2000.2594

122. Davita PWC, Pujiastuti H. Anallisis kemampuan pemecahan masalah matematika ditinjau dari gender. *Kreano J Mat Kreatif Inov.* (2020) 11:110–7. doi: 10.15294/kreano.v11i1.23601

123. Kalily S, Juhaevah F. Analisis Kemampuan Berpikir Kritis Siswa Kelas X SMA dalam Menyelesaikan Masalah Identitas Trigonometri Ditinjau Dari Gender. *J Mat Dan Pembelajaran*. (2018) 6:111–26.

124. Nisbett RE, Peng K, Choi I, Norenzayan A. Culture and systems of thought: holistic versus analytic cognition. *Psychol Rev*. (2001) 108:291–310. doi: 10.1037/0033-295X.108.2.291

125. Ji L, Vaughan-Johnston TI, Zhang Z, Jacobson JA, Zhang N, Huang X. Contextual and cultural differences in positive thinking. *J Cross Cult Psychol.* (2021) 52:449–67. doi: 10.1177/00220221211020442

TYPE Original Research
PUBLISHED 02 February 2023
DOI 10.3389/fpsyt.2023.1040861

Check for updates

OPEN ACCESS

EDITED BY
Jianzhong Su,
University of Texas at Arlington, United States

REVIEWED BY
Yuanxi Li,
East China University of Science and Technology, China
Songting Li,
Shanghai Jiao Tong University, China

*CORRESPONDENCE
Lisa Wagels
✉ lwagels@ukaachen.de

SPECIALTY SECTION
This article was submitted to Computational Psychiatry, a section of the journal Frontiers in Psychiatry

RECEIVED 09 September 2022
ACCEPTED 16 January 2023
PUBLISHED 02 February 2023

CITATION
Shymanskaya A, Kohn N, Habel U and Wagels L (2023) Brain network changes in adult victims of violence.
Front. Psychiatry 14:1040861.
doi: 10.3389/fpsyt.2023.1040861

Brain network changes in adult victims of violence

Aliaksandra Shymanskaya[1,2], Nils Kohn[3], Ute Habel[1,2] and Lisa Wagels[1,2]*

[1]Department of Psychiatry, Psychotherapy, and Psychosomatics, Faculty of Medicine, RWTH Aachen University, Aachen, Germany, [2]JARA-BRAIN Institute Brain Structure and Function, INM-10, Institute of Neuroscience and Medicine, Jülich Research Centre, Jülich, Germany, [3]Donders Institute for Brain, Cognition, and Behavior, Radboud University Medical Center, Nijmengen, Netherlands

Introduction: Stressful experiences such as violence can affect mental health severely. The effects are associated with changes in structural and functional brain networks. The current study aimed to investigate brain network changes in four large-scale brain networks, the default mode network, the salience network, the fronto-parietal network, and the dorsal attention network in self-identified victims of violence and controls who did not identify themselves as victims.

Materials and methods: The control group ($n = 32$) was matched to the victim group ($n = 32$) by age, gender, and primary psychiatric disorder. Sparse inverse covariance maps were derived from functional resting-state measurements and from T1 weighted structural data for both groups.

Results: Our data underlined that mostly the salience network was affected in the sample of self-identified victims. In self-identified victims with a current psychiatric diagnosis, the dorsal attention network was mostly affected underlining the potential role of psychopathological alterations on attention-related processes.

Conclusion: The results showed that individuals who identify themselves as victim demonstrated significant differences in all considered networks, both within- and between-network.

KEYWORDS

victims of violence, neuroimaging, structural covariance, functional connectivity, partial correlation, sparse inverse covariance

1. Introduction

The link between victimization and poor mental health has been recognized in many studies (1–3). Severe forms of victimization include physical and sexual violence. Additionally, different forms of abuse such as threat, stalking, blackmailing are often experiences as severe harm to an individual's life (4, 5) including consequences such as depression, anxiety, suicidal ideation, and panic attacks. The self-identification as a victim thereby does not necessarily agree with external labeling (1) and the same event might be perceived very differently between individuals. Importantly, even if not ensured by external sources, self-perceived victimization is stressful and associated with negative consequences such as self-blame, loneliness, anxiety, and low self-worth (6). Further studies suggest that executive functioning is reduced in individuals who have experienced violence during early childhood or adolescence (7, 8). Given the severe and often protracted effects of perceived victimization, it is important to determine how this can lead to mental problems. In this respect, the brain plays an important role. However, while victimization represents a severe risk factor for mental disorders, only little is known with respect to victimization as a trans-diagnostic risk factor on the neural level.

Previous studies have investigated if the exposure to violence can affect brain morphology and brain function. Such associations between violence exposures and brain structural changes have been investigated for gray (GM) and white matter (WM). In GM, changes in volume (9–13), cortical thickness (CT) (12–16), surface area (12–14, 16), and local gyrification (14) were observed in connection to the experience of childhood neglect and abuse. Mostly GM and CT reductions were observed in victims of violence [e.g., (11–14)]. Nevertheless, increases were reported in female survivors of intimate partner violence (10). Additionally, GM volume reductions in the prefrontal cortex were reported (17), and these findings were recently confirmed in a trans-diagnostic sample of adult participants who reported childhood maltreatment. Another line of studies investigated neural changes in association to combat-related trauma (9, 18, 19). Interestingly these studies showed that combat exposure related volume reductions were distinct from reductions related to a PTSD and depression diagnosis (20, 21). In sum, structural abnormalities were observed in both cortical and subcortical regions in different samples in all tissues, although some findings (22) argue against a strong association of WM changes and the experience of violence. Furthermore, a recent study pointed to alterations in brain organization (23) based on the covariance of GM volume between selected areas in victims versus controls. Studies that explicitly focus on the subjective self-identification as a victim including a broad definition of violence in this field are lacking.

Changes in brain activity also have been associated with the exposure to violence (24–28). The majority of fMRI studies in different populations that had been exposed to violence showed deviations in activation during cognitive or emotional tasks [for a review see (29)], and functional connectivity alterations occurred during emotion provoking tasks (26, 30, 31) and resting state fMRI (32). Functional differences between survivors of intimate partner violence (IPV) with a PTSD diagnosis and a non-traumatized group were reported in the anterior insula, which is the hub of the salience network (26). Furthermore, decreased connectivity among the anterior insula, amygdala, and anterior cingulate cortex (ACC), was reported for IPV related PTSD during a face-match task (31). Moreover, painful stimulation led to an elevated activation of the right middle insula and the right dorsolateral prefrontal cortex in IPV survivors with PTSD (33). Potential PTSD specific influence may be expected here, since a decrease in subjective pain intensity ratings over time was accompanied by attenuation of activation within the right anterior insula, which at the same time was associated with avoidance symptoms of PTSD.

Previous results in survivors of violence were often specific to a certain age group or a specific type of violence. Many studies have focused on physical, sexual or emotional abuse experienced during childhood (34), underlining the role of the hippocampus and amygdala (17, 35–37). Furthermore, in populations that experienced violence in early childhood, physical forms of violence seem to be associated more strongly with changes in amygdala and anterior cingulate cortex while emotional abuse may result in changes related to reward and mood processing circuits (38). Other findings may even suggest differences in the brain networks of individuals exposed to emotion abuse versus neglect (39). While some studies have successfully shown changes in brain activation for specific victimized populations, brain changes have–to our knowledge–not been studied in transdiagnostic samples of individuals who identified themselves as victim including a broad definition of victimization.

The existing literature demonstrates the necessity to study the relationship between brain modulations and subjective victimization as a trans-diagnostic phenomenon, thereby enabling the identification of neural changes independent of a mental health diagnosis. Additionally, specific types of experienced violence have mostly been investigated in specific groups, for example combat related exposure in males and intimate partner violence in females. To our knowledge, currently there is no study that included male and female adults identifying themselves as victims independent of the type of violence, or age of the individual. Furthermore, only a few studies have so far investigated large-scale network changes simultaneously on a structural and functional level. Specifically, changes in the default mode network (DMN), the fronto-parietal network (FPN), and the salience network (SN) as well as the dorsal attention network (DAN) (40–42) have been proposed as prominent characteristics of psychiatric disorders and as markers of exposure to violence. Thus, studying changes in these networks and their association with previous victimization may support the identification of neural risk factors for mental health issues, independent of a specific diagnosis.

The current study aimed to identify differences in structural and functional covariance in the DMN, FPN, SN, and DAN in two different groups: The first group (V) was characterized by self-identification as victim of violence; the second group (NV) was matched to the V group by age, gender and the primary psychiatric diagnosis. Our study did not exclude participants based on the psychiatric diagnosis and represented therefore a more realistic clinical population, which enabled the investigation of structural and functional brain network connectivity. To investigate structural and functional organization differences, we focused on pre-determined regions of interest (ROIs) in the DMN, FPN, SN, and DAN, and analyzed group differences in between and within network covariance patterns of both function and structure.

We expected to discover differences in structural and functional organization of the four large-scale brain networks between V and NV, independently of any psychiatric diagnosis. Similar changes in covariance in SN and DMN were expected in the V group. We assumed, that presence of a psychiatric diagnosis played an additional important role in the difference between V and NV. Therefore, we performed a diagnosis-specific explorative analysis. As a secondary hypothesis, we assumed that the group of V with a present acute psychiatric diagnosis (V_{D+}) differed from the NV with a present acute psychiatric diagnosis (NV_{D+}), and a differing structural and functional covariance pattern would be observed as compared to the trans-diagnostic consideration. As a third hypothesis, we assumed, that V and NV would differ in their psychopathology, which in turn would correlate with the differences in the structural and functional covariance.

2. Materials and methods

2.1. Sample

The sample included two groups of adults of which the first group had subjectively experienced violence (V), while the matched control group had not experienced violence before (NV). Inclusion criteria for both groups were: (i) age between 18 and 60 years, (ii) right-handedness, (iii) MRI suitability, and (iv) absence of any neurological diseases. Specific inclusion criterion for the V group was the prior experience of at least one type of the following forms of violence. The experience of violence was verified by a screening instrument and a detailed qualitative interview which were developed within the "Gender Violence" project (43, 44). The definition of violence used applied to this screening instrument and the interview based on the WHO standards (45) defining physical, emotional, and sexual violence. Orienting to previous studies, and because it is a frequent precursor or other forms of violence during intimate partnerships (46), economic violence (financial abuse) was added as a further category in the screening. Physical forms included all forms of body attacks such as hitting, kicking, shaking, spitting; sexual forms include all sexual acts without agreement such as coercion, sexual assault or rape; emotional forms included permanent insults, humiliation, bullying, stalking, threat; economic forms included robbery, passing of salary, prohibition to fulfill basic needs. The NV group included only individuals who negated any prior experience of violence at a primary screening and did not identify themselves as victimized. The V group was recruited from the participant pool of a large study in which detailed semi-structured interviews about the experience of violence were performed. Within this larger study, participants in the V group were recruited on the one hand in cooperation with an intervention center against domestic violence in Aachen, Germany ("Frauen helfen Frauen e.V.") who asked individuals with experiences of violence if they would be willing to participate in the study. Participation in the study was voluntary and completely independent of any further consultation. On the other hand, we distributed flyers describing different forms of violence, the study aims and contact points for individuals seeking help in all departments in the university hospital Aachen, including the emergency department and the psychiatric department. Individuals who self-identified themselves as victims of violence and wanted to participate could notify study personal *via* phone or email. Flyers were also distributed at other public places offering consultation or therapy to potential victimized individuals such as ambulant therapists. For individuals that were in addition to study participation or independent of study participation seeking help and that were not supported otherwise a team of trained experts and psychologists offered consultation as part of the project. The fMRI study only included individuals that had undergone the qualitative interview in the main arm of the study and a further screening concerning MRI criteria if participants were interested in taking part in the fMRI study. 33.3% of all recruited participants took part in the fMRI study as well. The NV group was directly recruited *via* flyers and at the university hospital RWTH Aachen, specifically the Department of Psychiatry, Psychotherapy and Psychosomatics. All participants gave their written informed consent to participate in the study and received a compensatory payment of 85 Euros. Included participants additionally underwent the Mini-International Neuropsychiatric Interview [MINI; (47)], which allowed us to match both groups for age, sex, and MINI diagnoses. Overall, the V group included 49 subjects and the NV group 41 individuals of which 25 in the V group and 20 in the NV group had any kind of psychiatric diagnosis.

2.2. Study protocol

The study protocol was approved by the internal Ethics Committee of the RWTH Aachen University and thus complied with the ethical principles stated in the Declaration of Helsinki. The complete study procedure consisted of an initial resting state fMRI scan, a social stress paradigm, an emotion induction paradigm, a second resting state scan, an anatomical scan, neuropsychological tests, and several self-report questionnaires.

Besides the behavioral variables, the present investigation focused on the anatomical scan and the first resting state scan. Imaging data were acquired on a whole-body Siemens 3T Trio scanner (Siemens AG; Erlangen, Germany) equipped with 12 channel head coil, located at the RWTH Aachen University hospital in Germany, whereas some subjects were measured after the scanner upgrade to Prisma. During the resting state acquisition, participants were instructed to relax and lie still with eyes opened, focusing a fixation cross presented on a black screen. Afterward, all participants assured that they had not fallen asleep.

In order to test if groups differed with regard to psychopathology severity, stress coping and neuropsychological functioning, after the MRI procedure, we quantified (i) the strength of depressive symptoms through the Beck Depression Inventory [BDI; (48)], (ii) state and trait anxiety scores through the State-Trait Anxiety Inventory [STAI; (49)], and (iii) information on stress exposure and stress symptoms through the Stress and Coping Inventory [SCI; (50)]. In the V group, we also measured perceived distress caused by violence experiences through the Impact of Event Scale [IES; (51)]. Neuropsychological tests included the digit span [ZNS, forward and backward; Hamburg Wechsler Intelligence test (HAWIE-R); (52)], the verbal fluency test [VLT; (53)], a measure for verbal intelligence [Mehrfach Wortschatztest version B, MWT_B; (54)] and a test for shared attention and executive functions/cognitive flexibility [Trail making test, TMT; (55)]. From the introduced neuropsychological tests, descriptive variables were derived: TMT comprised the difference between the acquired TMT version A and version B; VLT_1 represented the total fluency performance (i.e., phonemic fluency und semantic fluency), while VLT_2 represented switching (i.e., phonemic switching und semantic switching), and HAWIE-R (ZNS) represented the sum of the forward and backward digit-span tests.

Several participants could not be included in our analyses due to the following reasons: (i) missing anatomical ($n = 4$) or any resting state scans ($n = 7$) due to technical problems, (ii) incomplete coverage of the whole brain during structural scan ($n = 10$), (iii) influence of alcohol ($n = 1$), (iv) sudden nausea ($n = 1$), and (v) lack of credibility of statements due to several contradictions ($n = 1$). Thus, the final sample consisted of 64 participants, comprising 32 participants who experienced violence, and 32 controls. The number of V, who suffered from a current psychiatric diagnosis (V_{D+}), was 25, and the number of V without a current diagnosis (V_{D-}) was 7. The number of NV, who suffered from a current psychiatric diagnosis (NV_{D+}), was 20, and the number of NV without a diagnosis (NV_{D-}) was 12.

2.3. Voxel based morphometry

To investigate structural differences between both groups, we acquired a T1-weighed image for each participant using an MPRAGE sequence (TR = 2,300 ms, TE = 3.03 ms, flip angle = 9°, FOV = 256 × 256 mm^2, 176 sagittal slices, voxel size = 1 × 1 × 1 mm^3). Structural imaging data were preprocessed using the Computational Anatomy toolbox (CAT 12[1]). First, each scan was manually reoriented to the intercommisural plane. After correction for inhomogeneities in field intensity, affine and non-linear normalization to MNI standard space was applied using the DARTEL default template within a unified segmentation model (56). Then, images were segmented into GM, WM, and cerebrospinal fluid. Additionally, the GM volumes were scaled by the amount of contraction applied during the preceding normalization. This modulation with Jacobian determinants ensured that the total volume of GM corresponded to that of the original images. Finally, the modulated GM segments were smoothed using a Gaussian kernel of 8 mm FHWM which was suggested to improve the morphometric examination of smaller and larger brain regions (57, 58). A subsequent homogeneity check did not identify any outliers. The ensuing voxel-based morphometry data were used to examine covariance in GM volumes in the sample.

2.4. Functional resting state

To compare brain function between V and NV, 250 functional images for each participant were acquired using a EPI sequence (TR = 1,600 ms, TE = 30 ms, flip angle = 67°, FOV = 192 × 192 mm^2, matrix size = 64 × 64, 26 transversal slices, voxel size = 3 × 3 × 4.2 mm^3, acquisition order = interleaved ascending). Functional imaging data were preprocessed using the functional connectivity toolbox (CONN 18a[2]). Initially, the first four scans of each participant were discarded to allow for magnetic field saturation. Then, the individual resting state time series were preprocessed according to the following steps: (i) realignment and unwarping, (ii) slice-time correction, (iii) outlier detection [97th percentiles using Artifact Detection Toolbox (ART)], (iv) segmentation and spatial normalization to MNI standard space, and (v) smoothing (Gaussian kernel of 8 mm FWHM). Subsequently, the pre-processed time series were denoised to account for potential confounding effects of (i) 6 motion parameters, (ii) their derivatives, (iii) squares of the 6 motion parameters and their derivatives, (iv) mean CSF and WM signal (v) outlier regressors from ART. Additionally, quadratic detrending and despiking before regression were applied. We did not use global signal regression. Furthermore, the time-series were band-pass filtered to retain signals between 0.01 and 0.08 Hz. This frequency range likely represented neural signal and was less susceptible to physiological noise (59, 60). The resulting resting state time series were used to investigate functional connectivity in the sample.

2.5. ROI definition

We were interested in how covariance within and between for major networks differed between V and NV. For that aim, the functional connectivity toolbox CONN was used (61). CONN's standard network atlas was based on an independent component analysis of the functional resting state data of a large sample of healthy adults (61, 62). Although variances in the brain structure are expected in healthy controls and patient groups, applying the atlas information based on healthy adults for the investigation of patient groups is considered valid because previous studies have shown differences in the DNM, SN, and FPN based on different whole brain nodes and seeds [for a meta-analysis see Koch et al. (63)] suggesting robust group differences in these networks despite of potential structural differences. The atlas provides an established brain parcellation that divided the DMN, SN, DAN, and FPN into 19 spatially distinct network nodes, which were parts of the brain networks (**Figure 1**). The DMN covered the medial prefrontal cortex (MPFC), the bilateral lateral parietal cortex (LPCs), and the precuneus (PCUN). The SN included the anterior cingulate cortex (ACC) as well as the bilateral anterior insula (AIs), the rostral prefrontal cortex (RPFCs), and the supramarginal gyrus (SMGs). The DAN consisted bilaterally of frontal eye fields (FEFs) and the intraparietal sulci (IPSs). The FPN comprised both the right and left lateral prefrontal cortex (LPFCs) and the posterior parietal cortex (PPCs). The 19 investigated network nodes served as ROIs and were used to extract structural and functional brain information from individuals in the V and NV groups. For each participant, brain data was averaged across all voxels belonging to a particular ROI. This yielded individual average GM volumes, average GM density and average functional resting state time series for each ROI. The extracted GM volumes, densities and time series were z-standardized individually. This z-standardization mainly served two purposes in the following analyses: (i) to ensure the comparability of ROIs, and (ii) to enable the interpretation of covariance measures as correlation (= normalized covariance). To avoid potential confounding effects in the brain data, we accounted for sex, age, MINI diagnosis, antidepressants, and the number of other psychotropic drugs. In the structural analyses, we also accounted for total intracranial volume. Numerical confounds were z-standardized, while scale confounds were dummy encoded. Deconfounding on the group level was performed for time series in CONN, while for GM volumes and densities it was done with NiftiMapsMasker from nilearn package (64). The extracted network information served as input for the estimation of structural covariance and functional connectivity matrices in each group.

2.6. Sparse inverse covariance

To estimate covariance differences between the groups V/NV and V_{D+}/NV_{D+}, partial correlations between the 19 nodes were calculated on the group level. Group level partial correlations (assuming the group of subjects underlies the same functional or VBM structure) were calculated using sparse inverse covariance estimation [covariance precision from GraphicalLassoCV, from Python sklearn (64)]. By accounting for the influence of other brain regions, *partial* correlations as compared to *full* correlations yield direct, unbiased relationships between two ROIs (65, 66). Partial correlations could be estimated by sparse inverse covariance (67–69). An L1 penalty automatically set less important entries in the connectivity matrix to zero which enabled a robust estimation also in smaller samples (66, 69). The sparsity degree was internally chosen *via* a 3-fold cross-validation that ascertained the generalizability of the model to new data (70, 71). For calculations we used Python 3.7, primarily using the neuroimaging package nilearn (72) and the machine learning package scikit-learn (64).

To investigate the differences in structural covariance between V and NV, we estimated gray matter volumes (GMV), gray matter

[1] www.neuro.uni-jena.de/cat

[2] www.nitrc.org/projects/conn

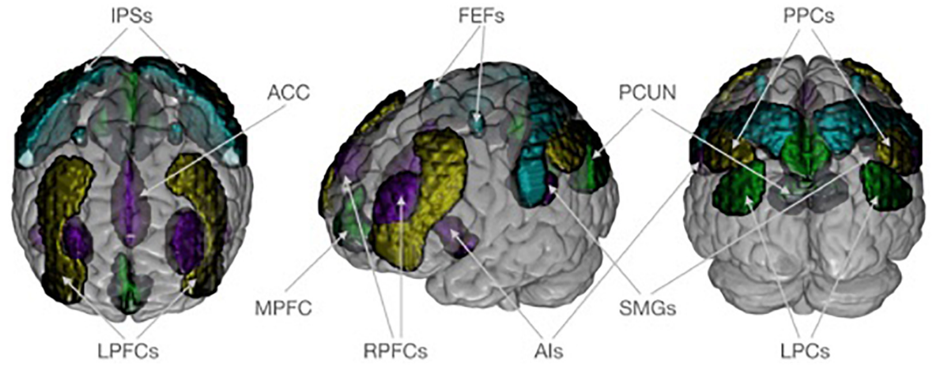


FIGURE 1
Neuroanatomical visualization of four target networks, according to CONN's standard atlas. The default mode network is shown in dark green and consists of the medial prefrontal cortex (MPFC), the bilateral lateral parietal cortex (LPCs), and the precuneus (PCUN). The saliency network (purple) includes the anterior cingulate cortex (ACC) as well as the bilateral anterior insula (AIs), rostral prefrontal cortex (RPFCs), and supramarginal gyrus (SMGs). The dorsal attention network (turquoise) consists of frontal eye field (FEFs) and intraparietal sulcus (IPSs) bilaterally. The frontoparietal network (yellow) comprises both the right and left lateral prefrontal cortex (LPFCs) and posterior parietal cortex (PPCs). Regions of interest are plotted on an MNI standard brain in anterior, lateral, and posterior view using MRIcroGL (https://www.nitrc.org/projects/mricrogl/).

densities (GMD) and gray matter masses (GMM) of ROIs, and performed independent *t*-tests, to see if there are significant group differences between the values of GMD, GMV or GMM. Further, we investigated partial correlations in the estimated GM parameters between the 19 ROIs on group level separately in the V and NV groups. For this purpose, we used sparse inverse covariance estimation to determine the partial correlation between brain volumes on group level and reported the differences in covariance between V and NV. A single value in each subject for a given GM volume in each ROI prohibited estimation of individual partial correlation matrices.

Afterward, we performed similar calculations for functional data, to investigate the differences in functional covariance between V and NV. An independent *t*-test probed for significant differences between the ROI time series of the V and NV groups. Again, sparse inverse covariance was used to generate partial correlations between brain volumes of each time series of resting state for each ROI. Analyses were performed separately for the V and NV groups. Described methods for GM variables and resting state are depicted in **Figure 2**.

The analysis steps described above for the determination of covariance differences, as well as statistical tests for group differences and brain-behavior associations, described in the following text, were repeated for a separate subgroup of subjects of the V and NV group with a current psychiatric diagnosis (D+). The subjects with a current diagnosis and an experience of violence (V_{D+}) were compared to those with a current diagnosis but without an experience of violence (NV_{D+}). These complementary calculations were performed to compare more homogeneous samples (see **Tables 2**, **3**). Due to a small number of participant without a psychiatric diagnosis, we did not perform sub-analyses in these small V and NV groups.

2.7. Group differences in covariance

Based on partial correlation maps, we examined group differences in covariance between ROIs. Negative values described a relative decrease in covariance, or cross-talk, between regions in the V group compared to the NV group. Complementarily, positive values described a relative increase of covariance, or higher level of cross-talk between the nodes. To inspect the observed structural and functional covariance patterns differences between victims and controls, we employed non-parametric test for mean differences (73, 74). To this end, we compared the data of V to the general distribution simulated by randomly drawing 10^4 bootstrap samples of the NV, with replacement. Thus, every bootstrapped subsample consisted of 41 subjects for the NV group, and 20 in the NV group had any kind of psychiatric diagnosis. Structural and functional correlation matrices of the victims were compared to the bootstrapped 99.999% population intervals of the controls, which corresponds to testing for significant differences at a corrected, two-sided alpha-level of 10^{-5}. The same analysis was performed for V_{D+} versus NV_{D+}.

2.8. Summary of covariance differences across imaging modalities

To determine the convergence of network-specific covariance differences between the pairs of 171 nodes, we summed up the findings by calculating the frequency of differences in covariance between V and NV in every node across all imaging modalities. This number thus represented the total number of group differences in a network, to which these nodes belonged. The same was done separately for the V_{D+} versus NV_{D+} comparison. Since four networks were investigated, ten possible combinations for within- and between-network covariance existed. Four of those described the *within-network* covariance (DMN-DMN, SN-SN, DAN-DAN, and FPN-FPN), and the rest described the *between-network* covariance (DMN-SN, DMN-DAN, DMN-FPN, SN-DAN, SN-FPN, DAN-FPN). Additionally, we aimed to investigate the importance of differing covariance for a specific single node. To this end, the frequency of each node being involved in a different covariance comparing V versus NV and V_{D+} versus NV_{D+} was recorded.

2.9. Brain-behavior associations

Finally, significant structural and functional network aberrations were probed for their association with specific behavioral and

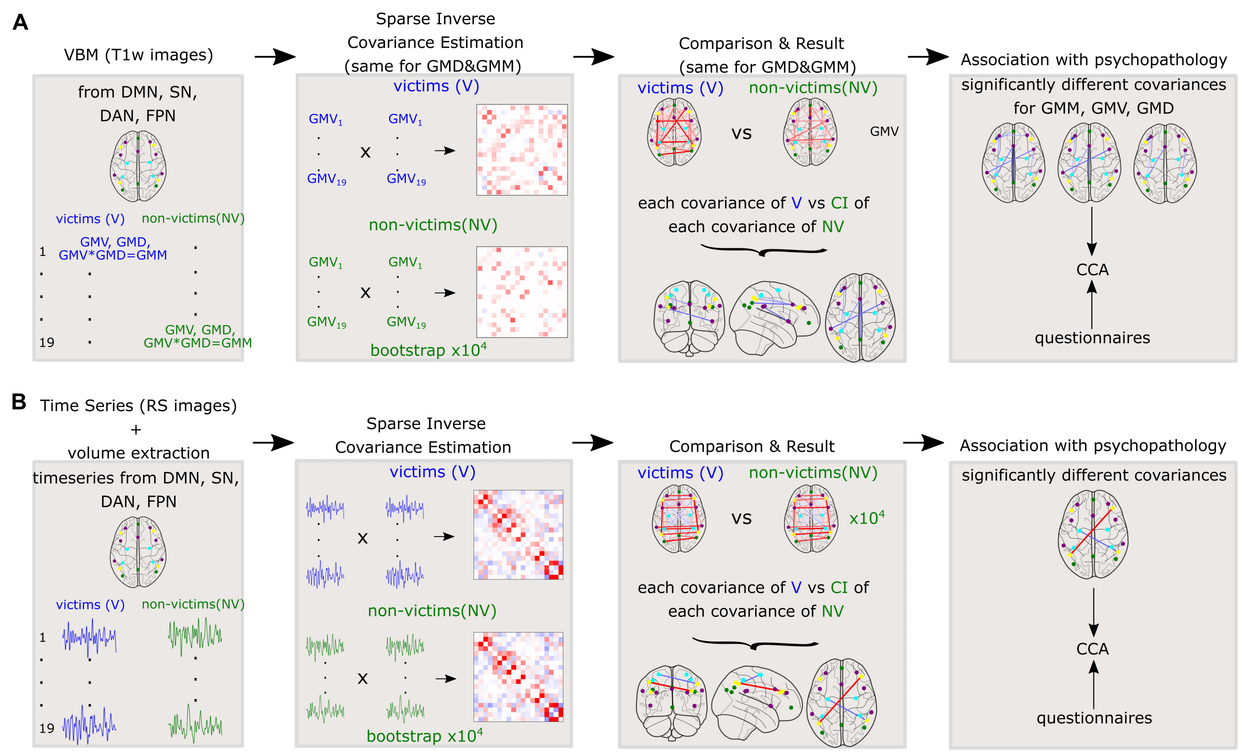

A
VBM (T1w images)
from DMN, SN, DAN, FPN
victims (V)
non-victims(NV)
GMV, GMD, GMV*GMD=GMM
1
19
Sparse Inverse Covariance Estimation (same for GMD&GMM)
victims (V)
GMV1
GMV19
x
non-victims(NV)
bootstrap x10^4
Comparison & Result (same for GMD&GMM)
vs
GMV
each covariance of V vs CI of each covariance of NV
Association with psychopathology
significantly different covariances for GMM, GMV, GMD
CCA
questionnaires
B
Time Series (RS images) + volume extraction
timeseries from DMN, SN, DAN, FPN
Sparse Inverse Covariance Estimation
Comparison & Result
x10^4
significantly different covariances

FIGURE 2
Methods applied to the whole cohort for **(A)**: Variables derived from gray matter (GM) such as gray matter volumes (GMV), gray matter densities (GMD), and gray matter masses (GMM) and **(B)**: Time series of resting state fMRI. Sparse inverse covariance was estimated between the 19 nodes on the group level. Further, the covariances were compared between V and NV. The differences in covariances were correlated with the questionnaires.

neuropsychological variables. We used canonical correlation analysis (CCA) which investigated internal relationships between two sets of variables by seeking maximal correlations between combinations of variables in both sets (75, 76). Thus, the aim of the CCA was to test if a significant amount of variance of structural/functional network aberrations and behavioral/neuropsychological variables across subjects could be explained by pairs of canonical variates (modes of co-variation), and to discover the internal relationships between the two sets (74). The latter was done by the calculation of the correlations between each variable and the corresponding canonical variates.

For functional data, we estimated covariance between the nodes, which demonstrated significant differences between the subject groups, for each participant on the individual level. For functional data, we estimated covariance for each participant on the individual level. Statistical significance of canonical correlations was determined sequentially with a Wilks' Lambda, using F-approximation (77). All p-values were Bonferroni-corrected to account for multiple comparisons and tested at a corrected alpha level of 0.05.

3. Results

3.1. Group characteristics

Within the V group experience of violence differed regarding the experiences type, length and age of exposure. In total, nine of 32 included participants in the V group experienced only emotional and economic violence and only one participant reported to have been exposed to economic violence as only form of violence. All other participants had experienced physical or sexual violence including multiple forms. Only 5 patients reported to have experienced physical violence without any other form of violence and only one participant reported to have been exposed to sexual violence only. Overall 22 participants reported to have been exposed to several forms of violence, while 10 reported only one form of violence. The estimated duration participants in the V group were exposed to one or more types of violence repeatedly was 8.3 years with only three individuals being exposed to violence (physical) only a single time. While the estimated mean duration did not differ significantly between individuals who had experienced physical or sexual violence (among others) and those who did not [$t(30) = 1.19$, $p = 1.22$, Table 1]. However, age of the first exposure was significantly lower individuals who experiences physical or sexual violence compared to those who did not [$t(30) = 2.02$, $p = 0.035$, Table 1]. Comparing the mean scores of the IES subscales intrusion, avoidance and hyperarousal of individuals who experienced (among others) physical or sexual violence in contrast (23) to those who did not report any of these forms (9) showed no significant differences in any scale [intrusion: $t(28) = 1.36$, $p = 0.092$; avoidance: $t(28) = 0.31$, $p = 0.389$; hyperarousal: $t(28) = 0.58$; $p = 0.285$]. Two participants (experiencing physical forms of violence) did not want to answer the IES, which is why the mean scores are reported only for a group of 30 participants.

As shown by non-significant group differences, we were able to successfully match self-identified victims and the control group in age, sex, and MINI diagnoses (Table 2). The groups

TABLE 1 Mean and standard deviation of the subscales on the impact of events scale, the estimated duration of exposure to violence and the age at the first exposure to violence in the V group contrasting individuals who had experiences physical or sexual forms of violence to those who exclusively experienced other forms.

	Physical or sexual violence		No physical or sexual violence	
	M	SD	*M*	SD
Intrusion	18.33	9.096	13.33	9.552
Avoidance	20.95	9.870	19.78	8.700
Hyperarousal	16,15	10,520	13.78	9.615
Duration of exposure	3400.49	2996.61	2095.33	2124.99
Age at first exposure	11.83	9.62	24.33	17.54

TABLE 2 Sample characteristics for victims of violence (V) and non-victims (NV): Binary variables (sex, MINI diagnosis, and antidepressants): Statistical comparison of groups performed *via* chi-square test of independence; Continuous variables: mean ± standard deviation, statistical comparison of groups performed *via* independent two-sample *t*-test for normally distributed features [STAI trait, TMT, VLT$_2$, HAWIE-R (ZNS)] and Mann Whitney *U*-test for non-parametric cases.

	V group (*N* = 32)	NV group (*N* = 32)	Group differences (*p*-values)
Sex	20♀ and 12♂	18♀ and 14♂	1.00
Age	33.3 ± 10.1	32.5 ± 11.7	1.00
BDI	15.5 ± 11.2	8.9 ± 10.4	0.033
STAI trait	46.1 ± 12.5	35.5 ± 16.3	0.060
SCI stress exposure	64.0 ± 19.8	47.9 ± 22.3	0.035
MINI diagnosis	78.1%	62.5%	1.00
Antidepressants	38.0%	25.0%	1.00
Number of other psychotropic drugs	0.2 ± 0.4	0.1 ± 0.4	0.588
Number of violence experiences	1.9 ± 0.9	—	
Childhood violence	1.3 ± 0.5	—	
TMT	−20.4 ± 12.5	−14.1 ± 10.6	0.232
VLT$_1$	35.2 ± 8.6	39.4 ± 6.5	0.076
VLT$_2$	29.5 ± 6.2	33.3 ± 4.8	0.101
HAWIE-R (ZNS)	14.9 ± 3.8	15.3 ± 3.8	1.00
MWT_B	28.3 ± 6.0	30.9 ± 3.2	0.113

Normality distribution was tested using Shapiro Wilk test. The equality of variance was tested with the Levene test. *P*-values were Bonferroni corrected at the significance level of 0.05.

did not significantly differ in any neuropsychological variable or questionnaire.

Comparison of groups only including participants with a current psychiatric diagnosis (V_{D+} and NV_{D+}) were provided in Table 3. The groups did not significantly differ in any neuropsychological variable or questionnaire. Furthermore, no significant differences were discovered neither between GMM, GMV or GMD nor between the time series in ROIs of the V and NV groups and the V_{D+} and NV_{D+} groups.

3.2. Structural covariance in brain networks in all subjects, independent on the psychiatric diagnosis

Based on the structural covariance matrix, and after performing previously described bootstrapping to identify differences in covariance between V and NV, we identified 9 out of 171 node pairs, that demonstrated differences in GMM covariance between groups at a corrected alpha level of 10^{-4} (Figure 3A). Significantly less covariance between regions in V was observed in all significantly different covariance measures. *Within-network* disturbances emerged only in the SN and constituted 33% of all detected aberrations, demonstrating less covariance and therefore lower homogeneity in the structural organization of V as compared to NV. *Between-network-wise*, FPN, DAN, SN, and DMN revealed aberrations in half of their nodes. On the other hand, GMD (Figure 3C) demonstrated 4 aberrant connections, all of which overlapped with GMM covariance differences. These consistent aberrations were observed in SN/DAN and SN/FPN. GMV (Figure 3E) demonstrated 5 aberrant connections, which showed reduced covariance in V, same as in GMM. These aberrations were observed in GMM, and occurred within SN, in SN/FPN, SN/DMN, and SN/DAN.

3.3. Functional covariance in brain networks in all subjects, independent of psychiatric diagnosis

The comparison of functional covariance matrices yielded 2 out of 171 functional connections that significantly differed between groups at a corrected alpha-level of 10^{-4} (Figure 4A). Significantly *higher* covariance between regions in V was observed *within* the FPN network. *Between-network-wise*, less covariance was observed in V in FPN/DAN covariance.

We furthermore investigated differences in structural and functional covariance between groups with a current MINI diagnosis (V_{D+} and NV_{D+}).

3.4. Structural covariance in brain networks in subjects with a current psychiatric diagnosis

We identified 13 out of 171 pairs of nodes (Figure 3B) that differed significantly in GMM between V_{D+} and NV_{D+}. Only one node (DMN/DAN) showed a slight overexpression in V_{D+}, as compared to NV_{D+}. Lower covariance between regions in V_{D+} was observed in the remaining 12 connections. Within-network disturbances emerged only in the SN. Between-network-wise, the aberrations were observed in DMN, FPN and DAN across all modalities. GMV (Figure 3D) demonstrated 10 aberrant connections, which showed less covariance in V_{D+}. Within-network covariance aberrations were observed in SN, and in DAN. In contrast to the whole sample GMV analysis (V/NV), group differences in the V_{D+}/NV_{D+} sample demonstrated a decrease in DAN covariance with DMN and FPN. Furthermore, GMD demonstrated 9 aberrant connections, with significant decrease in covariance in DAN/DMN and DAN/FPN, and with within-network covariance aberrations in SN (Figure 3F).

TABLE 3 Sample characteristics for victims of violence (V_{D+}) and non-victims (NV_{D+}) with a current diagnosis: Binary variables (sex, MINI diagnosis, and antidepressants): Statistical comparison of groups performed *via* chi-square test of independence; Continuous variables: mean ± standard deviation, statistical comparison of groups performed *via* independent two-sample *t*-test for normally distributed features [SCI stress exposure, TMT, VLT$_2$, HAWIE-R (ZNS)] and Mann Whitney *U*-test for non-parametric cases (the rest).

	V_{D+} group ($N = 25$)	NV_{D+} group ($N = 20$)	Group differences (p-values)
Sex	14♀ and 11♂	12♀ and 8♂	1.00
Age	33.2 ± 9.9	32.8 ± 12.0	1.00
BDI	17.3 ± 11.3	11.7 ± 10.9	1.00
STAI trait	47.0 ± 12.2	36.9 ± 18.9	0.263
SCI stress exposure	68.9 ± 18.8	53.3 ± 25.1	1.00
Antidepressants	36.0%	35.0%	1.00
Number of other psychotropic drugs	0.2 ± 0.4	0.2 ± 0.5	1.00
Number of violence experiences	2.0 ± 1.0	—	
Childhood violence	1.3 ± 0.5	—	
TMT	−21.6 ± 12.2	−12.5 ± 10.5	0.077
VLT$_1$	33.6 ± 8.3	40.9 ± 7.2	0.022
VLT$_2$	28.4 ± 6.2	33.5 ± 4.7	0.055
HAWIE-R (ZNS)	14.6 ± 3.9	14.7 ± 3.9	1.00
MWT_B	27.7 ± 6.5	30.7 ± 3.1	0.190

Normality distribution was tested using Shapiro Wilk test. The equality of variance was tested with the Levene test. *P*-values were Bonferroni corrected at the significance level of 0.05.

3.5. Functional covariance in brain networks in subjects with a current psychiatric diagnosis

The comparison of functional covariance matrices yielded 4 out of 171 functional connections, that significantly differed between the V_{D+} and NV_{D+} groups at a corrected alpha-level of 10^{-4} (Figure 4B). Covariance in V_{D+} differed from NV_{D+} in the same four nodes as in the V versus NV group, and in the three additional nodes. In these nodes, SN/DMN and SN/FPN demonstrated higher, and DAN/DMN lower covariance in V_{D+} as compared to NV_{D+}. No *within-network* differences were observed.

3.6. Across-modality covariance differences

The summaries of structural and functional covariance differences between the two V and NV groups across all modalities were depicted in Figure 5A for V versus NV, and in Figure 5B for V_{D+} versus NV_{D+}. The histograms demonstrated that covariance in V_{D+} was to a higher degree different from NV_{D+}, than in the analogous comparison of the V versus NV. Specifically, this meant that the networks in V_{D+} were less covariant between each other, than in NV_{D+}. In particular, regarding the V_{D+} versus NV_{D+} comparison, the DAN was associated with the majority of within- and between-network covariance differences (33% of all differences), followed by the SN (28%). In contrast, the main sample (V versus NV) showed mostly differences in the SN (49% of all differences), followed by the FPN (26%). Upon characterizing the affected network nodes on the individual level (Figure 6), we demonstrated that the DMN and DAN nodes were affected to a lower degree in the main sample than in the D + sample (10 versus 14% for DMN and 20 versus 27% for DAN). On the other hand, SN and FPN were affected more in the full sample (40 versus 32% for SN and 30 versus 27% for FPN).

3.7. Association with psychopathology

Finally, we examined if the observed group differences in structural and functional covariance were related to psychopathological symptoms (STAI trait, BDI, SCI stress exposure), as well as to neuropsychological functions [VFT$_1$ with one, and VFT$_2$ with two categories, TMT, MWT and HAWIE-R (ZNS)] using CCA. In the full sample (V versus NV), the analysis revealed a single highly significant CCA mode that related brain connectomes to subject measures ($r = 0.94$, $p = 0.008$). We observed that 94% of the variation in brain connectomes was explained by the variation in questionnaires. Since only the first CCA mode was significant, the first canonical variate for brain measures (CCX_1) was plotted against the first canonical variate for the questionnaires (CCY_1) in the scatter plot (Figure 7). These correlations between each variable and the corresponding canonical variate were used to interpret the first CCA mode, and the correlations with the correlation over $r > 0.2$ were provided in Table 4. The contributions of the variables to the CCA modes were also demonstrated in Figure 7. All correlations between the first canonical variable for brain, and the brain measures were uniformly large, and were represented by all included measures of DMN and SN. Among the psychopathological symptoms and neuropsychology variables, STAI contributed to CCY_1 to the highest proportion. Thus, CCY_1 can be considered as an anxiety measure. Thus, due to the significance of the CCA decomposition, CCX_1 and CCY_1 demonstrated high correlation, and uncovered dependence between anxiety traits and a linear combination of structural brain measures of SN and DMN. However, the CCX_1 and CCY_1 did not differ significantly between V and NV. Thus, the latent variables did not reflect the victimization status.

In the D+ groups, the number of subjects was not sufficient to estimate the modes of variance reliably for both RS and GM brain measures. However, based on the previous analysis of the full sample, we considered for the CCA analysis only GM brain measures. The analysis revealed a single highly significant CCA mode that related brain connectomes to subject measures ($r = 0.99$, $p < 10^{-5}$). Again, the first canonical variate for brain measures (CCX_1) was plotted against the first canonical variate for the questionnaires (CCY_1) in the scatter plot (Figure 8). The highest correlations between each variable and the corresponding canonical variate were provided in Table 5. The contributions of the variables to the CCA modes were also demonstrated in Figure 8. The correlations between the first canonical variable for brain, and the brain measures were uniformly large, and were again represented by the measures of DMN and SN. Therefore, the canonical variate CCX_1 could again be considered as an overall measure across all brain measures. On the other hand, TMT and ZNS contributed to CCY_1 to the highest proportion. Thus, the dependence between TMT and ZNS, and a linear combination of structural brain measures of SN and DMN, was discovered.

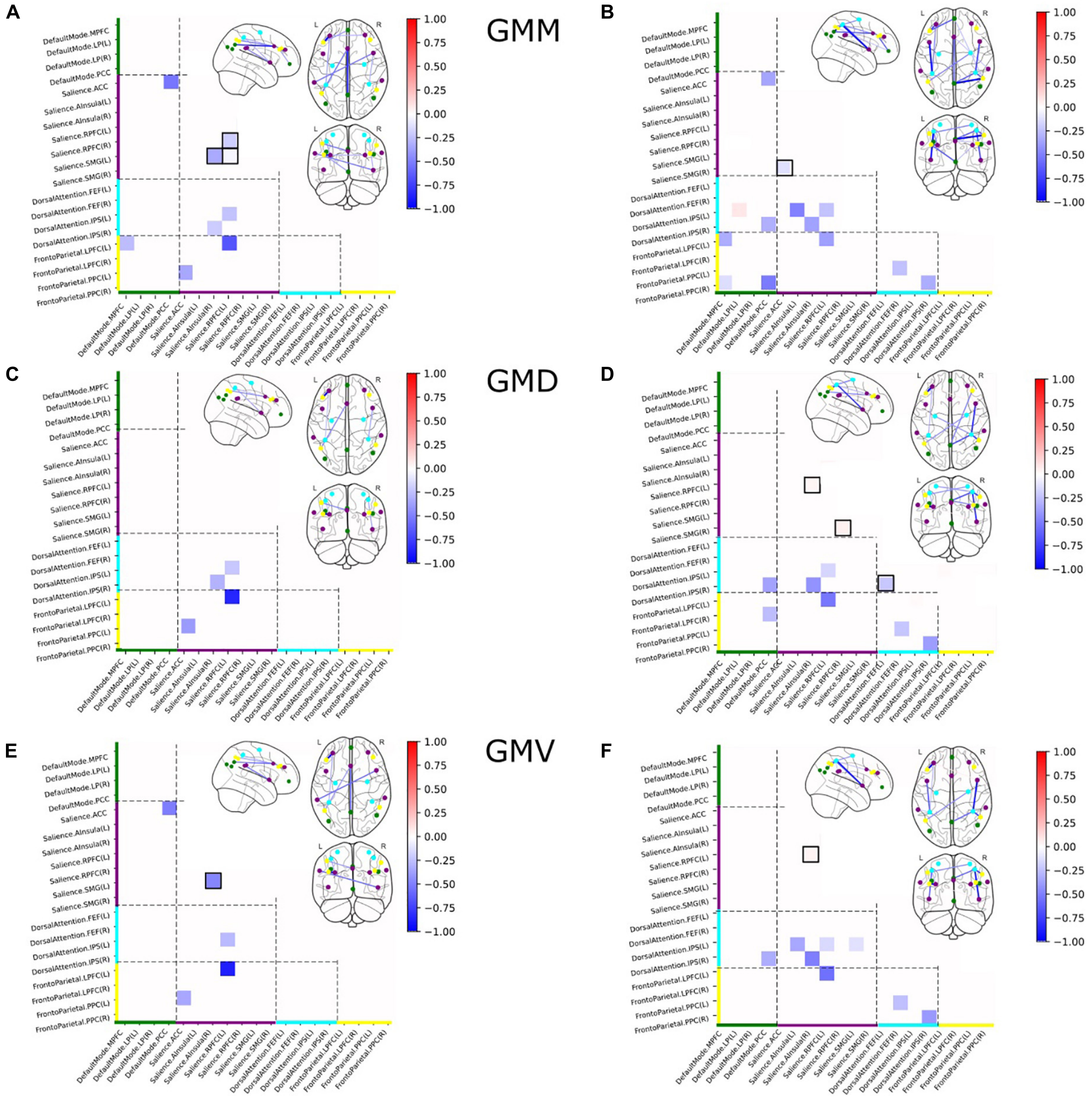


FIGURE 3
Group differences in structural covariance of gray matter masses (GMM) **(A)**, gray matter densities (GMD) **(C)**, and gray matter volumes (GMV) **(E)** within and between four major brain networks for the full sample, and group differences in structural covariance of GMM **(B)**, GMV **(D)**, and GMD **(F)** within and between four major brain networks for the subsample of subjects with a current psychiatric diagnosis. Squares indicate significant differences in partial correlations between V and NV. Colors on the axes and of the nodes correlated with the networks: DMN–green, SN–purple, DAN–cyan, FPN–yellow. Nodes with within-network differences were highlighted with black squares.

4. Discussion

The current study aimed to identify differences in structural and functional covariance in the DNM, FPN, SN, and DAN in a trans-diagnostic sample of individuals who identified themselves as victims of violence. This sample was compared to individuals who did not indicate any prior experience of violence but had a similar history of mental disorders. To further limit the influence of different psychopathologies on the differences in network covariance in both groups, two comparisons were made: a comparison of victims and non-victims in the whole sample (V versus NV), and in a subsample with the present psychiatric diagnosis (D+ group: V_{D+} versus NV_{D+}). Applying multiple comparisons correction, the only differences between V and NV was discovered in BDI and SCI stress exposure, both higher in V. On the other hand, the only difference between V_{D+} versus NV_{D+} was discovered in VLT_1, which was higher in NV. These differences may be linked to the exposure to violence indirectly as the patients in this study who mostly had experienced violence over a long time and with multiple incidents may have had an increased severity of emotional and cognitive symptoms compared to other patients who may not have had any traumatic experiences.

While no differences between the V and NV group nor between V_{D+} and NV_{D+} group were observed neither in GMM/GMV/GMD nor in the time series in ROIs, the relative organization of the brain seemed to be different between groups. Specifically, differences in

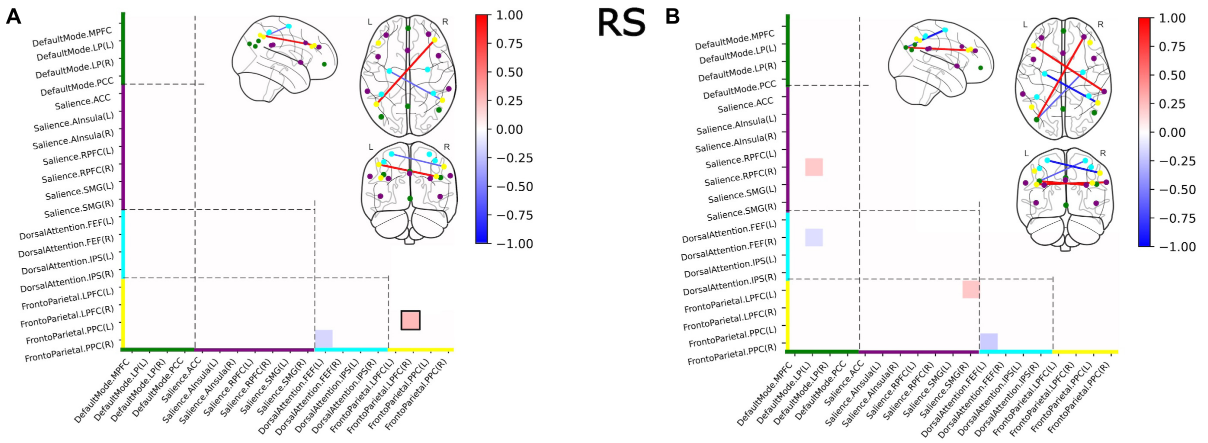


FIGURE 4
Group differences in functional covariance of RS within and between four major brain networks for the full sample **(A)** and for the subsample of subjects with a current psychiatric diagnosis **(B)**. Squares indicate significant differences in partial correlations between victims and controls. Colors on the axes and of the nodes correlated with the networks: DMN–green, SN–purple, DAN–cyan, FPN–yellow. Nodes with within-network differences were highlighted with black squares.

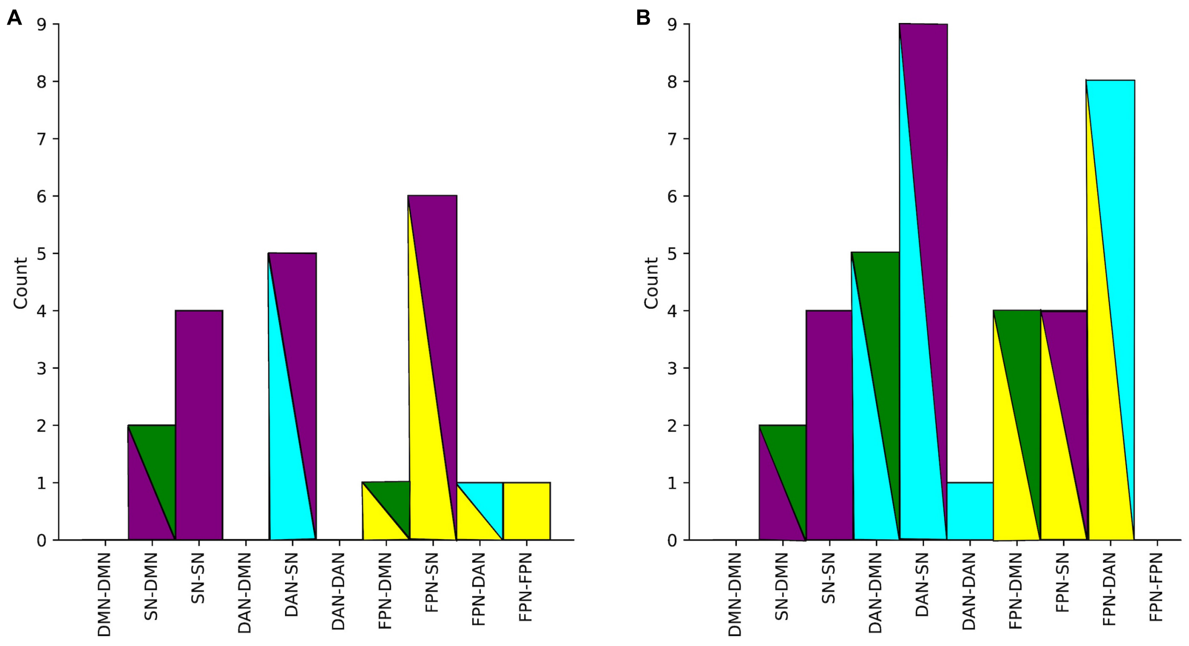


FIGURE 5
(A) Number of all significant group differences in covariance between networks in the main sample. **(B)** Differences in covariance between networks in D+ sample (subjects with a psychiatric diagnosis). Colors represent networks: DMN–green, SN–purple, DAN–cyan, FPN–yellow.

structural and functional covariance within and between the four selected networks were discovered. Sparse inverse covariance of the GM parameters and RS time series between the regions showed both positive and negative partial covariance differences within and between networks in both the full sample comparisons, and in the D+ sample comparisons.

Differences in the covariance in all four investigated networks, detected in V versus NV, may reflect organizational differences in the brain of victimized individuals related to specific characteristics of the group. However, we did not find any correlation between observed neural differences and the self-reports. There may be different explanations: on the one hand, self-reports may have not reliably reflected well-being and psychopathological symptoms due to the influence of self-perceptual abilities and social desirability. On the other hand, differences in network organization may have had heterogeneous sources or were related to further variables not specifically assessed in this study. While single values of psychopathological symptoms and neuropsychology did not differ in comparisons of V versus NV and V_{D+} versus NV_{D+}, CCA analysis discovered significant CCA modes in both cases. This way, the questionnaires of the full sample, mainly represented by anxiety, related to the brain measures of DMN and SN. On the other hand, neural measures of the DMN and SN in the D+ sample explained the variability in shared attention and executive functions/cognitive

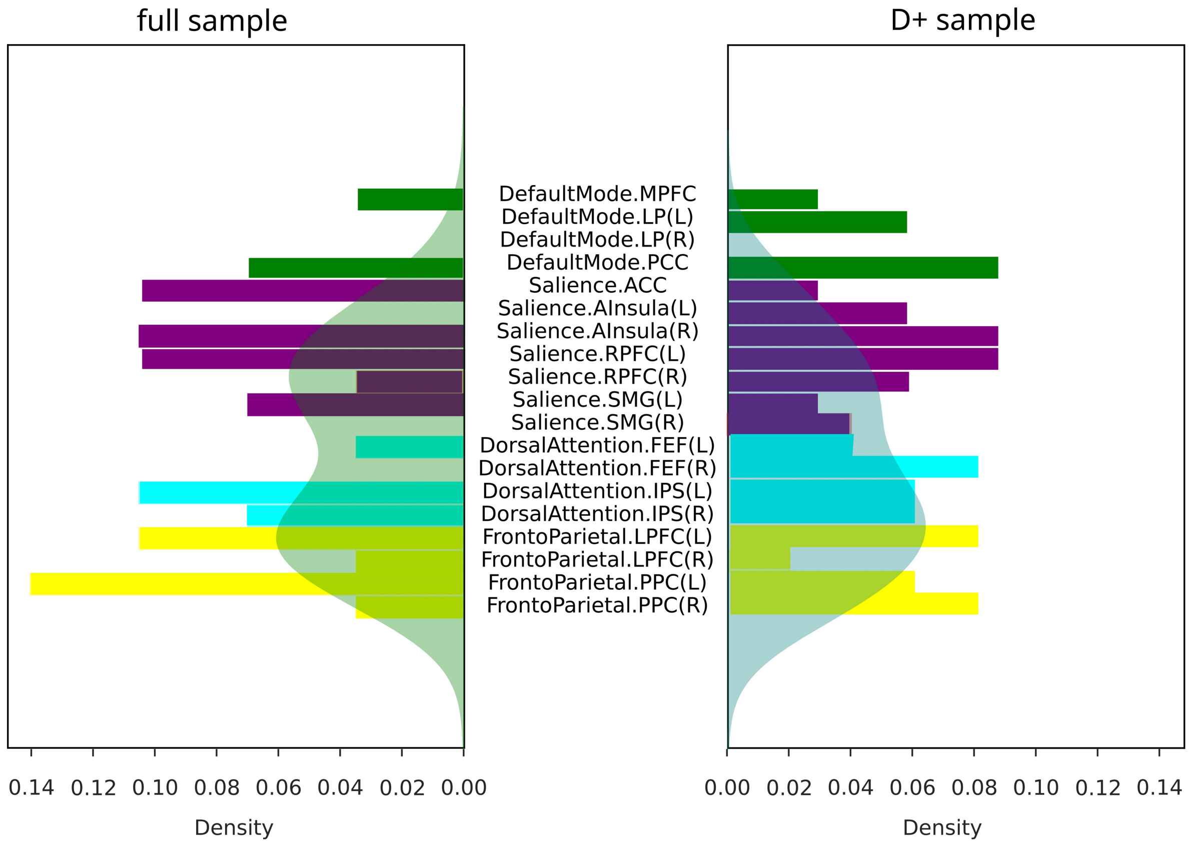


FIGURE 6

Covariance differences across modalities in all subjects (full sample for V/NV comparison, and D+ sample for V_{D+}/NV_{D+} comparison). Colors represent networks: DMN–green, SN–purple, DAN–cyan, FPN–yellow. Shaded areas represented density functions of the histograms and highlighted differences in frequency of node involvement.

flexibility as well as working memory, based on the uncovered single significant CCA mode. Evidence was found for the relationship between working memory and executive functioning, which might point to the common executive attention construct (78). While no direct evidence was found, the strongly affected covariance of the DAN might underlie the observed CCA modality in the D+ sample. Nevertheless, the latent variables did not reflect the victimization status, since the CCX_1 and CCY_1 did not differ between V_{D+} and NV_{D+}.

Structural and functional differences did not show a large overlap, which may further support the heterogeneous sources of variance in the self-identified victims and non-victims. In patients with major depression that experienced childhood trauma disturbances in functional brain networks similar to those investigated in our study have been associated with trauma severity (79). Higher childhood trauma severity moreover predicted symptoms of anxiety which may show some similarity to the association of the anxiety related component and covariance measures of gray matter in SN and DMN. In addition to factors that directly relate to the exposure of and severity to violence, cultural influences, personality, genetics, and for the patients in both groups also access and success of mental health treatment, may contribute to the reorganization of brain structure and function. These different sources of variance may impact structure and function differently thereby concealing or enhancing organizational changes in structure or function. Our results underline what has been summarized in a systematic review on subtypes of violence and associated functional and structural alterations; deviations occur in different brain regions not only depending on the subtype of violence but also with regard to structure and function activity and connectivity or integrity (39). The biological pathway

TABLE 4 Correlation between canonical correlation analysis (CCA) variates and variables in the full sample.

Brain measure variable	Correlation with CCX_1
GMD DMN.LP (R)	−0.21
GMD DMN.PCC	−0.30
GMM SN.AInsula (L)	−0.20
GMM SN.RPFC (L)	−0.23
GMV DMN.LP (R)	−0.20
GMV SN.AInsula (L)	−0.21
Questionnaire variables	**Correlation with CCY_1**
BDI	−0.27
STAI trait	−0.83
SCI stress exposure	−0.39
HAWIE-R (ZNS)	−0.28
VFT_1	−0.25
VFT_2	−0.29

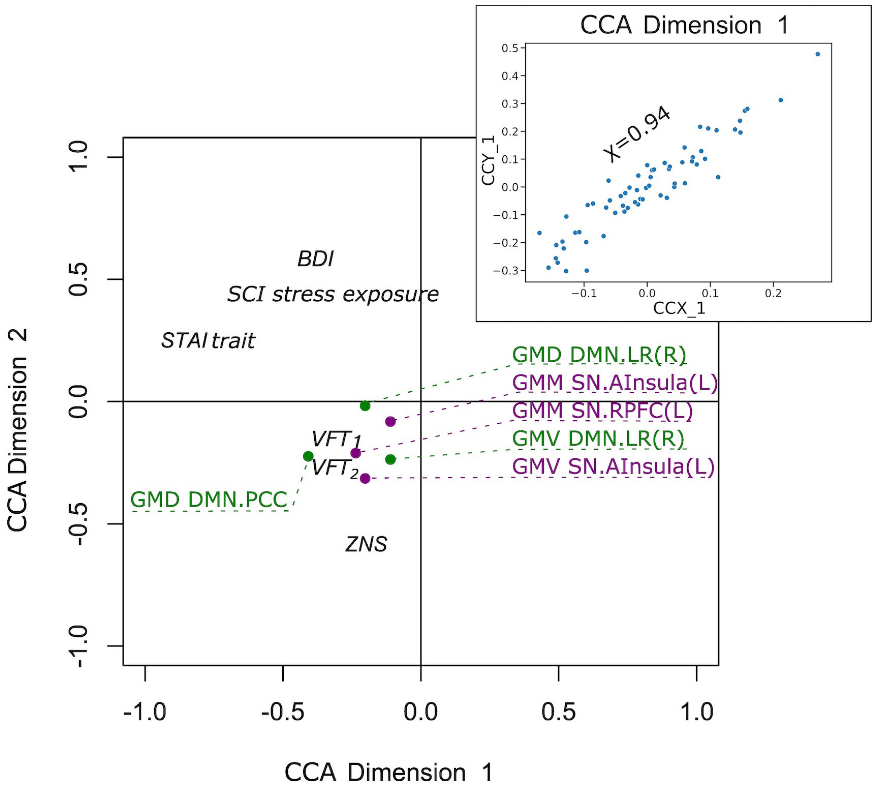


FIGURE 7
Correlation of the first canonical correlation analysis (CCA) mode variables for the full sample (upper right), and contributions of the variables to the first and second CCA dimensions.

the authors suggest may be one reason for such differences. Early-life stress (exposure to violence) is expected to affect brain organization which then may result in functional network changes either directly or rather indirectly accompanying pathology. Instead of originally affected stress-related brain regions, regions that are associated with other cognitive processes affected by dysfunctional stress systems may show functional disturbances in later life. Observing such a mismatch of structural and functional covariance measures may thus support the independence or asymmetry of structural and functional network organization. The SN was affected in both the full sample and the subjects with a current diagnosis (D+), with it being the most affected network in the full sample. Structurally, in the full sample and in the D+ sample, the SN demonstrated *reduced* within- and between-network covariance. Functionally, however, *no differences* relative to NV were observed in the full sample, while in V_{D+} SN demonstrated *increased* covariance in SN/DMN and SN/FPN. Thus, V_{D+} in our study demonstrated functional hyperconnectivity of SN, which was not observed in victims if the group also included healthy, potentially more resilient individuals. It could be therefore hypothesized, that especially those victims with a present diagnosis exhibited a functionally disturbed SN. Nevertheless, the victimized group proved to suffer from structural aberrations in the SN covariance. In healthy populations, the SN has been recognized as necessary for the efficient regulation of activity in the DMN. Thus, the failure of this regulation would lead to inefficient cognitive control and weaker performance on cognitive control tasks (80). Correspondingly, although in our sample patients in the V and NV group were mostly not free of a psychiatric diagnosis, possibly especially the victim group suffered from loss of control. Similarly, in Bogliacino et al. (81), a reduction of cognitive control in victims of urban violence and warfare was reported. Furthermore, differences involving SN in the D+ sample were most frequent in DAN/SN, while in the full sample this was the case for the FPN/SN covariance. The latter network communication seems to be responsible for the externally directed cognition (80). Finding altered communication between FPN/SN in the whole V sample may thus underline that these regions are affected independent of the severity or violence or the mental health consequences.

Node connections of the DMN also demonstrated major differences between V and NV in both structure and function. The DMN in healthy subjects is responsible for a self-referential introspective state (82). Structurally, both the full sample and the D+ sample demonstrated *reduced* between-network covariance in DMN/FPN. DMN/FPN connection was shown to be responsible for introspective processes and executive function (83). Thus, the reduced structural covariance between DMN and FPN might point to possible *decreased* introspective processes, often observed in victims elsewhere (83). While this would have to be confirmed in future studies, such a deficit might be more pronounced in participants in the V_{D+} group due to an active psychiatric diagnosis. Functionally, the inverse interplay between psychopathology and large-scale brain networks of the DMN/FPN has been demonstrated before (41) for a number of psychiatric diagnoses. This might explain the observed higher number of differences in DMN/FPN structural covariance as compared to the full sample in our study. Functionally, however, we could not support this finding. While no differences were discovered in the full sample, V_{D+} demonstrated reduced covariance in DMN/DAN (i.e., the connection DMN.PCC/DAN.IPS), and

TABLE 5 Correlation between canonical correlation analysis (CCA) variates and variables for the sample of subjects with a current psychiatric diagnosis.

Brain measure variable	Correlation with CCX_1
GMV DMN.LP (R)	0.42
GMV SN.AInsula (R)	0.43
GMD SN.AInsula (R)	0.39
GMM SN.AInsula (L)	0.41
GMM SN.SMG (L)	0.43
Questionnaire variables	**Correlation with CCY_1**
BDI	−0.36
MWT_B	0.26
TMT	0.68
HAWIE-R (ZNS)	0.71
VFT_1	0.44
VFT_2	0.57

increased covariance in DMN/SN. DMN/DAN covariance has been related to perceptual attention in healthy populations (83), while in anxiety and PTSD patients, functional covariance impairments were observed in dorsolateral prefrontal cortex (84). While DMN/DAN covariance seemed to be similar between V and NV, we cannot make a definitive statement in this regard in the D+ sample due to opposite covariance between different nodes. More research into the single nodes is required at this point.

The DAN is engaged during externally directed attentional tasks (85). In V and V_{D+}, the DAN demonstrated *reduced* within- and between-network covariance both in structure and function. Additionally, to the DAN/SN and DAN/DMN differences in covariance, described above, D+ sample demonstrated a higher number of differences in structural covariance in DAN/FPN. FPN regulates DAN in accordance with goals and task demands, and it is involved in the regulation of perceptual attention (83). Recent data showed negatively associated network connectivity between DAN and FPN in subjects with depression, anxiety and suicidality (41). Thus, based on our findings and in line with others, the higher proportion of DAN/FPN covariance differences in D+ sample might be a sign of the less efficient attention processes as compared to the full sample.

In interpreting these findings, several limitations have to be taken into account. First, only selected networks were investigated, therefore differences in other networks cannot be excluded. These networks were anatomically defined which may introduce a larger bias than extracting data-driven time series as in other studies (86). The victims were self-identified victims of violence, which is a highly subjective measure, and it cannot be quantified, since no correlation between aberrant network nodes and behavioral variables were discovered. Nevertheless, it is important to investigate neural alterations related to the subjective perception as this perception may be strongly connected to mental health problems (6). Despite of attempts to account for a large heterogeneity with regard to

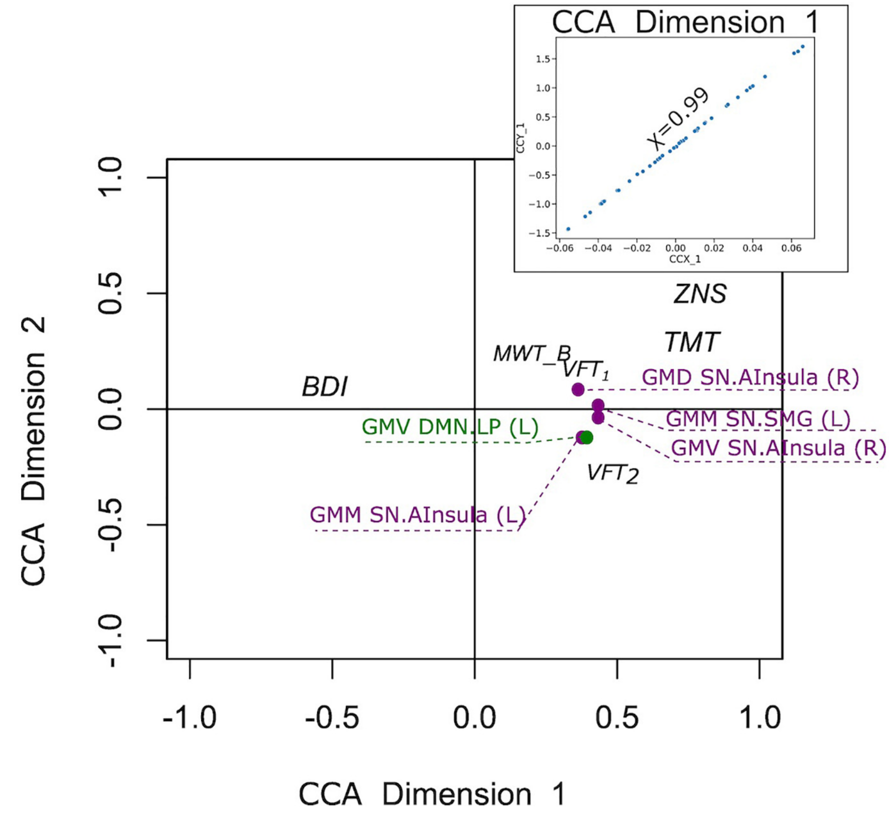


FIGURE 8
Correlation of the first canonical correlation analysis (CCA) mode variables for the sample of subjects with a current psychiatric diagnosis (upper right), and contributions of the variables to the first and second CCA dimensions.

psychopathology by matching both groups with regard to the primary psychiatric diagnosis, the total sample was too small to test different subgroups with specific disorders and subgroups without any mental disorder. In addition, the heterogeneity of the type of violence individuals were exposed to was large and the size of the sample did not allow us to test in network changes may differ depending on specific forms of violence such as (exclusively social) or non-social forms of violence. As pointed out in a recent review (38) physical and sexual violence in early childhood may seems to be associated with higher risks of PTSD and personality disorders while emotional violence more often associated with developing major depression. Animal models of physical versus non-physical abuse even suggest that brain circuit changes associated with abuse may differ. The current results, referring to the in changes of brain connectivity across all different types of violence may therefore conceal more specific changes associated with physical or non-physical violence. Further research in single nodes and in subgroups must be performed, while the study sample is to be extended. Finally, the upgrade of the scanner to Prisma while the study was carried may have introduced data variance which can reduce the classification accuracy in the data as shown in projects applying classifiers on fMRI data in multi-side projects (87).

In a nutshell, differences in functional and structural covariance between self-identified victims and people who never experienced violence or did not identify themselves as victims were observed, with a primary role of the SN. In the group with heightened pathologies and various mental disorders, most differences between victims and non-victims occurred in DAN. When the sample was controlled for psychiatric disorders, less covariance differences were observed, indicating that a major part of the network variance may reflect differences in the pathological status of two groups.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving human participants were reviewed and approved by Internal Ethics Committee of the RWTH Aachen University. The patients/participants provided their written informed consent to participate in this study.

Author contributions

LW, NK, and UH designed the study. LW collected the data. AS and LW preprocessed and analyzed the data and wrote the manuscript. All authors provided feedback and approved the final manuscript version.

Funding

This work was funded by the State of North Rhine-Westphalia (NRW, Germany), the European Union through the “NRW Ziel2 Program” as a part of the European Fund for Regional Development, and by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)–269953372/GRK2150.

Acknowledgments

This work was supported by the Brain Imaging Facility of the Interdisciplinary Centre for Clinical Research (IZKF) Aachen within the Faculty of Medicine at RWTH Aachen University.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher’s note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Arseneault L, Bowes L, Shakoor S. Bullying victimization in youths and mental health problems: ‘much ado about nothing’? *Psychol Med.* (2010) 40:717–29. doi: 10.1017/S0033291709991383

2. Hedtke KA, Ruggiero KJ, Fitzgerald MM, Zinzow HM, Saunders BE, Resnick HS, et al. A longitudinal investigation of interpersonal violence in relation to mental health and substance use. *J Consult Clin Psychol.* (2008) 76:633–47. doi: 10.1037/0022-006X.76.4.633

3. Lacey KK, McPherson MD, Samuel PS, Powell Sears K, Head D. The impact of different types of intimate partner violence on the mental and physical health of women in different ethnic groups. *J Interpers Violence.* (2013) 28:359–85. doi: 10.1177/0886260512454743

4. Dokkedahl S, Kok RN, Murphy S, Kristensen TR, Bech-Hansen D, Elklit A. The psychological subtype of intimate partner violence and its effect on mental health: protocol for a systematic review and meta-analysis. *Syst Rev.* (2019) 8:198. doi: 10.1186/s13643-019-1118-1

5. Stevens F, Nurse JRC, Arief B. Cyber stalking, cyber harassment, and adult mental health: a systematic review. *Cyberpsychol Behav Soc Netw.* (2021) 24:367–76. doi: 10.1089/cyber.2020.0253

6. Graham S, Juvonen J. Self-blame and peer victimization in middle school: an attributional analysis. *Dev Psychol.* (1998) 34:587–99. doi: 10.1037/0012-1649.34.3.587

7. Daly BP, Hildenbrand AK, Turner E, Berkowitz S, Tarazi RA. Executive functioning among college students with and without history of childhood maltreatment. *J Aggress Maltreat Trauma.* (2017) 26:717–35. doi: 10.1080/10926771.2017.1317685

8. Li Y, Dong F, Cao F, Cui N, Li J, Long Z. Poly-victimization and executive functions in junior college students. *Scand J Psychol.* (2013) 54:485–92. doi: 10.1111/sjop.12083

9. Clausen AN, Billinger SA, Sisante J-FV, Suzuki H, Aupperle RL. Preliminary evidence for the impact of combat experiences on gray matter volume of the posterior insula. *Front Psychol.* (2017) 0:2151. doi: 10.3389/FPSYG.2017.02151

10. Daugherty J, Verdejo-Román J, Pérez-García M, Hidalgo-Ruzzante N. Structural brain alterations in female survivors of intimate partner violence. *J Interpers Violence.* (2020) 37:1–34. doi: 10.1177/0886260520959621

11. Fonzo GA, Flagan TM, Sullivan S, Allard CB, Grimes EM, Simmons AN, et al. Neural functional and structural correlates of childhood maltreatment in women with intimate-partner violence-related posttraumatic stress disorder. *Psychiatry Res Neuroimaging.* (2013) 211:93–103. doi: 10.1016/J.PSCYCHRESNS.2012.08.006

12. Lim L, Hart H, Mehta M, Worker A, Simmons A, Mirza K, et al. Grey matter volume and thickness abnormalities in young people with a history of childhood abuse. *Psychol Med.* (2018) 48:1034–46. doi: 10.1017/S0033291717002392

13. Price M, Albaugh M, Hahn S, Juliano AC, Fani N, Brier ZMF, et al. Examination of the association between exposure to childhood maltreatment and brain structure in young adults: a machine learning analysis. *Neuropsychopharmacology.* (2021) 46:1888–94. doi: 10.1038/s41386-021-00987-7

14. Kelly PA, Viding E, Wallace GL, Schaer M, De Brito SA, Robustelli B, et al. Cortical thickness, surface area, and gyrification abnormalities in children exposed to maltreatment: neural markers of vulnerability? *Biol Psychiatry.* (2013) 74:845–52. doi: 10.1016/J.BIOPSYCH.2013.06.020

15. Ross MC, Sartin-Tarm AS, Letkiewicz AM, Crombie KM, Cisler JM. Distinct cortical thickness correlates of early life trauma exposure and posttraumatic stress disorder are shared among adolescent and adult females with interpersonal violence exposure. *Neuropsychopharmacology.* (2020) 46:741–9. doi: 10.1038/s41386-020-00918-y

16. Tozzi L, Garczarek L, Janowitz D, Stein DJ, Wittfeld K, Dobrowolny H, et al. Interactive impact of childhood maltreatment, depression, and age on cortical brain structure: mega-analytic findings from a large multi-site cohort. *Psychol Med.* (2020) 50:1020–31. doi: 10.1017/S003329171900093X

17. Paquola C, Bennett MR, Lagopoulos J. Understanding heterogeneity in grey matter research of adults with childhood maltreatment–a meta-analysis and review. *Neurosci Biobehav Rev.* (2016) 69:299–312. doi: 10.1016/J.NEUBIOREV.2016.08.011

18. Averill LA, Abdallah CG, Pietrzak RH, Averill CL, Southwick SM, Krystal JH, et al. Combat Exposure severity is associated with reduced cortical thickness in combat veterans: a preliminary report. *Chron Stress.* (2017) 1. doi: 10.1177/2470547017724714

19. Bremner JD, Randall P, Scott TM, Bronen RA, Seibyl JP, Southwick SM, et al. MRI-based measurement of hippocampal volume in patients with combat-related posttraumatic stress disorder. *Am J Psychiatry.* (1995) 152:973. doi: 10.1176/AJP.152.7.973

20. Averill CL, Satodiya RM, Scott JC, Wrocklage KM, Schweinsburg B, Averill LA, et al. Posttraumatic stress disorder and depression symptom severities are differentially associated with hippocampal subfield volume loss in combat veterans. *Chron Stress.* (2017) 1:2017. doi: 10.1177/2470547017744538

21. Wrocklage KM, Averill LA, Cobb Scott J, Averill CL, Schweinsburg B, Trejo M, et al. Cortical thickness reduction in combat exposed U.S. veterans with and without PTSD. *Eur Neuropsychopharmacol.* (2017) 27:515–25. doi: 10.1016/J.EURONEURO.2017.02.010

22. Fennema-Notestine C, Stein MB, Kennedy CM, Archibald SL, Jernigan TL. Brain morphometry in female victims of intimate partner violence with and without posttraumatic stress disorder. *Biol Psychiatry.* (2002) 52:1089–101. doi: 10.1016/s0006-3223(02)01413-0

23. Roos A, Fouche J-P, Stein DJ. Brain network connectivity in women exposed to intimate partner violence: a graph theory analysis study. *Brain Imaging Behav.* (2017) 11:1629–39. doi: 10.1007/s11682-016-9644-0

24. Elton A, Tripathi SP, Mletzko T, Young J, Cisler JM, James GA, et al. Childhood maltreatment is associated with a sex-dependent functional reorganization of a brain inhibitory control network. *Hum Brain Mapp.* (2014) 35:1654–67. doi: 10.1002/HBM.22280

25. Ethridge P, Sandre A, Dirks MA, Weinberg A. Past-year relational victimization is associated with a blunted neural response to rewards in emerging adults. *Soc Cogn Affect Neurosci.* (2018) 13:1259–67. doi: 10.1093/SCAN/NSY091

26. Simmons AN, Paulus MP, Thorp SR, Matthews SC, Norman SB, Stein MB. Functional activation and neural networks in women with posttraumatic stress disorder related to intimate partner violence. *Biol Psychiatry.* (2008) 64:681–90. doi: 10.1016/J.BIOPSYCH.2008.05.027

27. Strigo IA, Simmons AN, Matthews SC, Grimes EM, Allard CB, Reinhardt LE, et al. Neural correlates of altered pain response in women with posttraumatic stress disorder from intimate partner violence. *Biol Psychiatry.* (2010) 68:442–50. doi: 10.1016/j.biopsych.2010.03.034

28. Weissman DG, Jenness JL, Colich NL, Miller AB, Sambrook KA, Sheridan MA, et al. Altered neural processing of threat-related information in children and adolescents exposed to violence: a transdiagnostic mechanism contributing to the emergence of psychopathology. *J Am Acad Child Adolesc Psychiatry.* (2020) 59:1274–84. doi: 10.1016/J.JAAC.2019.08.471

29. Hein TC, Monk CS. Research review: neural response to threat in children, adolescents, and adults after child maltreatment–a quantitative meta-analysis. *J Child Psychol Psychiatry.* (2017) 58:222–30.

30. Cisler JM, Scott Steele J, Smitherman S, Lenow JK, Kilts CD. Neural processing correlates of assaultive violence exposure and PTSD symptoms during implicit threat processing: a network-level analysis among adolescent girls. *Psychiatry Res Neuroimaging.* (2013) 214:238–46. doi: 10.1016/J.PSCYCHRESNS.2013.06.003

31. Fonzo GA, Simmons AN, Thorp SR, Norman SB, Paulus MP, Stein MB. Exaggerated and disconnected insular-amygdalar blood oxygenation level-dependent response to threat-related emotional faces in women with intimate-partner violence posttraumatic stress disorder. *Biol Psychiatry.* (2010) 68:433–41. doi: 10.1016/j.biopsych.2010.04.028

32. Boccadoro S, Siugzdaite R, Hudson AR, Maeyens L, Van Hamme C, Mueller SC. Women with early maltreatment experience show increased resting-state functional connectivity in the theory of mind (ToM) network. *Eur J Psychotraumatol.* (2019) 10:1647044. doi: 10.1080/20008198.2019.1647044

33. Strigo IA, Simmons AN, Matthews SC, Grimes EM, Allard CB, Reinhardt LE, et al. Neural correlates of altered pain response in women with posttraumatic stress disorder from intimate partner violence. *Biol Psychiatry.* (2010) 68:442–50.

34. McCrory EJ, Gerin MI, Viding E. Annual research review: childhood maltreatment, latent vulnerability and the shift to preventative psychiatry – the contribution of functional brain imaging. *J Child Psychol Psychiatry.* (2017) 58:338–57. doi: 10.1111/JCPP.12713

35. Lim L, Howells H, Radua J, Rubia K. Aberrant structural connectivity in childhood maltreatment: a meta-analysis. *Neurosci Biobehav Rev.* (2020) 116:406–14. doi: 10.1016/J.NEUBIOREV.2020.07.004

36. Yuan M, Rubin-Falcone H, Lin X, Rizk MM, Miller JM, Sublette ME, et al. Smaller left hippocampal subfield CA1 volume is associated with reported childhood physical and/or sexual abuse in major depression: a pilot study. *J Affect Disord.* (2020) 272:348–54. doi: 10.1016/J.JAD.2020.03.169

37. van Rooij SJH, Smith RD, Stenson AF, Ely TD, Yang X, Tottenham N, et al. Increased activation of the fear neurocircuitry in children exposed to violence. *Depress Anxiety.* (2020) 37:303–12. doi: 10.1002/DA.22994

38. Waters RC, Gould E. Early life adversity and neuropsychiatric disease: differential outcomes and translational relevance of rodent models. *Front Syst Neurosci.* (2022) 16:860847. doi: 10.3389/fnsys.2022.860847

39. Cassiers LL, Sabbe BG, Schmaal L, Veltman DJ, Penninx BW, Van Den Eede F. Structural and functional brain abnormalities associated with exposure to different childhood trauma subtypes: a systematic review of neuroimaging findings. *Front Psychiatry.* (2018) 9:329. doi: 10.3389/fpsyt.2018.00329

40. Sha Z, Wager TD, Mechelli A, He Y. Common dysfunction of large-scale neurocognitive networks across psychiatric disorders. *Biol Psychiatry.* (2019) 85:379–88.

41. Yu M, Linn KA, Shinohara RT, Oathes DJ, Cook PA, Duprat R, et al. Childhood trauma history is linked to abnormal brain connectivity in major depression. *Proc Natl Acad Sci U.S.A.* (2019) 116:8582–90. doi: 10.1073/pnas.1900801116

42. Lebois LA, Li M, Baker JT, Wolff JD, Wang D, Lambros AM, et al. Large-scale functional brain network architecture changes associated with trauma-related dissociation. *Am J Psychiatry.* (2021) 178:165–73.

43. Evler A, Scheller M, Wagels L, Bergs R, Clemens B, Kohn N, et al. Gender-inclusive care of victims of violence: the model project "gender Gewaltkonzept" at the university hospital Aachen. *Der Nervenarzt.* (2016) 87:746–52. doi: 10.1007/s00115-015-0024-6

44. Habel U, Wagels L, Ellendt S, Scheller M, Evler A, Bergs R, et al. Violence and health. Symptoms, consequences and treatment of victimized patients. *Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz.* (2016) 59:17–27. doi: 10.1007/s00103-015-2258-7

45. Garcia-Moreno C, Jansen HA, Ellsberg M, Heise L, Watts CH. Prevalence of intimate partner violence: findings from the WHO multi-country study on women's health and domestic violence. *Lancet.* (2006) 368:1260–9. doi: 10.1016/S0140-6736(06)69523-8

46. Hing N, O'Mullan C, Nuske E, Breen H, Mainey L, Taylor A, et al. Gambling-related intimate partner violence against women: a grounded theory model of individual and relationship determinants. *J Interpers Violence.* (2022) 37:N18639–65. doi: 10.1177/08862605211037425

47. Sheehan DV, Lecrubier Y, Sheehan KH, Amorim P, Janavs J, Weiller E, et al. The Mini-International Neuropsychiatric Interview (M.I.N.I.): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *J Clin Psychiatry.* (1998) 59(Suppl. 20):quiz34–57.

48. Beck AT, Ward CH, Mendelson M, Mock J, Erbaugh J. An inventory for measuring depression. *Arch General Psychiatry.* (1961) 4:561–71. doi: 10.1001/archpsyc.1961.01710120031004

49. Spielberger CD, Gorsuch RL, Lushene RE. *The State-Trait Anxiety Inventory (Test Manual).* Palo Alto, CA: Consulting Psychologists Press (1970).

50. Satow L. *Stress–und Coping-Inventar (SCI): Test–und Skalendokumentation.* (2012). Available online at: http://www.drsatow.de/ (accessed February 15, 2013).

51. Horowitz M, Wilner N, Alvarez W. Impact of event scale: a measure of participantive stress. *Psychosom Med.* (1979) 41:209–18. doi: 10.1097/00006842-197905000-00004

52. Erzberger CS, Engel RR. Zur äquivalenz der normen des wechsler-intelligenztests für erwachsene (WIE) mit denen des Hamburg-wechsler-intelligenztests für erwachsene – revision (HAWIE-R). *Z Neuropsychol.* (2010) 21:25–37. doi: 10.1024/1016-264X/a000002

53. Troyer AK, Moscovitch M, Winocur G. Clustering and switching as two components of verbal fluency: evidence from younger and older healthy adults. *Neuropsychology.* (1997) 11:138–46. doi: 10.1037/0894-4105.11.1.138

54. Lehrl S. *Mehrfachwahl-Wortschatz-Intelligenztest MWT-B. Manual zum MWTB.* Balingen: Spitta-Verl (1995).

55. Tischler L, Petermann F. Trail making test (TMT). *Z Psychiatrie Psychol Psychother.* (2010) 58:79–81. doi: 10.1024/1661-4747.a000009

56. Ashburner J, Friston KJ. Unified segmentation. *Neuroimage.* (2005) 26:839–51.

57. Honea R, Crow TJ, Passingham D, Mackay CE. Regional deficits in brain volume in schizophrenia: a meta-analysis of voxel-based morphometry studies. *Am J Psychiatry.* (2005) 162:2233–45. doi: 10.1176/appi.ajp.162.12.2233

58. White T, OLeary D, Magnotta V, Arndt S, Flaum M, Andreasen NC. Anatomic and functional variability: the effects of filter size in group fMRI data analysis. *Neuroimage.* (2001) 13:577–88. doi: 10.1006/nimg.2000.0716

59. Fox DF, Raichle ME. Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. *Nat Rev Neurosci.* (2007) 8:700–11. doi: 10.1038/nrn2201

60. Lu H, Zuo Y, Gu H, Waltz JA, Zhan W, Scholl CA, et al. Synchronized delta oscillations correlate with the resting-state functional MRI signal. *Proc Natl Acad Sci USA.* (2007) 104:18265–9. doi: 10.1073/pnas.0705791104

61. Whitfield-Gabrieli S, Nieto-Castanon A. Conn: a functional connectivity toolbox for correlated and anticorrelated brain networks. *Brain Connectiv.* (2012) 2:125–41. doi: 10.1089/brain.2012.0073

62. Rademacher J, Galaburda AM, Kennedy DN, Filipek PA, Caviness VS. Human cerebral cortex: localization, parcellation, and morphometry with magnetic resonance imaging. *J Cogn Neurosci.* (1992) 4:352–74. doi: 10.1162/jocn.1992.4.4.352

63. Koch SB, van Zuiden M, Nawijn L, Frijling JL, Veltman DJ, Olff M. Aberrant resting-state brain activity in posttraumatic stress disorder: a meta-analysis and systematic review. *Depress Anxiety.* (2016) 33:592–605. doi: 10.1002/da.22478

64. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res.* (2011) 12:2825–30.

65. Marrelec G, Horwitz B, Kim J, Pelegrini-Issac M, Benali H, Doyon J. Using partial correlation to enhance structural equation modeling of functional MRI data. *Magnet Reson Imaging.* (2007) 25:1181–9. doi: 10.1016/j.mri.2007.02.012

66. Varoquaux G, Craddock RC. Learning and comparing functional connectomes across participants. *Neuroimage.* (2013) 80:405–15.

67. Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics.* (2008) 9:432–41. doi: 10.1093/biostatistics/kxm045

68. Smith SM, Miller KL, Salimi-Khorshidi G, Webster M, Beckmann CF, Nichols TE, et al. Network modelling methods for FMRI. *Neuroimage.* (2011) 54:875–91.

69. Varoquaux G, Gramfort A, Poline J-B, Thirion B. Brain covariance selection: better individual functional connectivity models using population prior. *Proceedings of the Advances in Neural Information Processing Systems.* Vancouver, BC: (2010). p. 2334–42.

70. Stone M. Cross-validatory choice and assessment of statistical predictions. *J R Stat Soc Ser B (Methodol).* (1974) 36:111–47. doi: 10.1111/j.2517-6161.1974.tb00994.x

71. Stone M. Cross-validation: a review. *Statistics.* (1978) 9:127–39.

72. Abraham A, Pedregosa F, Eickenberg M, Gervais P, Mueller A, Kossaifi J, et al. Machine learning for neuroimaging with scikit-learn. *Front Neuroinformatics.* (2014) 8:14. doi: 10.3389/fninf.2014.00014

73. Miller KL, Alfaro-Almagro F, Bangerter NK, Thomas DL, Yacoub E, Xu J, et al. Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nat Neurosci.* (2016) 19:1523–36.

74. Smith SM, Nichols TE, Vidaurre D, Winkler AM, Behrens TEJ, Glasser MF, et al. A positive-negative mode of population covariation links brain connectivity, demographics and behavior. *Nat Neurosci.* (2015) 18:1565–7. doi: 10.1038/nn.4125

75. Wang H-T, Smallwood J, Mourao-Miranda J, Xia CH, Satterthwaite TD, Bassett DS, et al. Finding the needle in a high-dimensional haystack: canonical correlation analysis for neuroscientists. *Neuroimage.* (2020) 216:116745. doi: 10.1016/j.neuroimage.2020.116745

76. Zhuang X, Yang Z, Cordes D. A technical review of canonical correlation analysis for neuroscience applications. *Hum Brain Mapp.* (2020) 41:3807–33. doi: 10.1002/hbm.25090

77. Mardia KV, Kent JT, Bibby JM. *Multivariate Analysis*. Cambridge, MA: Academic Press (1979).

78. McCabe DP, Roediger HL, McDaniel MA, Balota DA, Hambrick DZ. The relationship between working memory capacity and executive functioning: evidence for a common executive attention construct. *Neuropsychology.* (2010) 24:222–43. doi: 10.1037/a0017619

79. Schirmer ST, Beckmann FE, Gruber H, Schlaaff K, Scheermann D, Seidenbecher S, et al. Decreased functional connectivity in patients with major depressive disorder and a history of childhood traumatization through experiences of abuse. *Behav Brain Res.* (2023) 437:114098. doi: 10.1016/j.bbr.2022.114098

80. Menon V, Uddin LQ. Saliency, switching, attention and control: a network model of insula function. *Brain Struct Funct.* (2010) 214:655–67. doi: 10.1007/s00429-010-0262-0

81. Bogliacino F, Grimalda G, Ortoleva P, Ring P. Exposure to and recall of violence reduce short-term memory and cognitive control. *Proc Natl Acad Sci USA.* (2017) 114:8505–10. doi: 10.1073/pnas.1704651114

82. Mak LE, Minuzzi L, MacQueen G, Hall G, Kennedy SH, Milev R. The default mode network in healthy individuals: a systematic review and meta-analysis. *Brain Connectiv.* (2017) 7:25–33. doi: 10.1089/brain.2016.0438

83. Dixon ML, De La Vega A, Mills C, Andrews-Hanna J, Spreng RN, Cole MW, et al. Heterogeneity within the frontoparietal control network and its relationship to the default and dorsal attention networks. *Proc Natl Acad Sci USA.* (2018) 115:E1598–607. doi: 10.1073/pnas.1715766115

84. Wen Z, Seo J, Pace-Schott EF, Milad MR. Abnormal dynamic functional connectivity during fear extinction learning in PTSD and anxiety disorders. *Mol Psychiatry.* (2022) 27:2216–24. doi: 10.1038/s41380-022-01462-5

85. Spreng RN, Shoemaker L, Turner GR. Executive functions and neurocognitive aging. In: Goldberg E editor. *Executive Functions in Health and Disease.* Amsterdam: Elsevier (2017). p. 169–96. doi: 10.1016/B978-0-12-803676-1.00008-8

86. Nicholson AA, Harricharan S, Densmore M, Neufeld RW, Ros T, McKinnon MC, et al. Classifying heterogeneous presentations of PTSD via the default mode, central executive, and salience networks with machine learning. *Neuroimage Clin.* (2020) 27:102262. doi: 10.1016/j.nicl.2020.102262

87. Kang L, Chen J, Huang J, Jiang J. Autism spectrum disorder recognition based on multi-view ensemble learning with multi-site fMRI. *Cogn Neurodyn.* (2022).

TYPE Review
PUBLISHED 07 February 2023
DOI 10.3389/fpsyt.2023.1092471

Understanding mental health through computers: An introduction to computational psychiatry

Juan Camilo Castro Martínez[1]* and
Hernando Santamaría-García[2,3,4]

[1]Departamento de Psiquiatría y Salud Mental, Facultad de Medicina, Pontificia Universidad Javeriana, Bogotá, Colombia, [2]Ph.D. Programa de Neurociencias, Departamento de Psiquiatría y Salud Mental, Pontificia Universidad Javeriana, Bogotá, Colombia, [3]Centro de Memoria y Cognición Intellectus, Hospital Universitario San Ignacio, Bogotá, Colombia, [4]Global Brain Health Institute, University of California, San Francisco – Trinity College Dublin, San Francisco, CA, United States

Check for updates

OPEN ACCESS

EDITED BY
Jianzhong Su,
The University of Texas at Arlington,
United States

REVIEWED BY
Xiaochuan Pan,
East China University of Science
and Technology, China
Kamran Afzali,
Université de Montréal, Canada

*CORRESPONDENCE
Juan Camilo Castro Martinez
✉ j-castrom@javeriana.edu.co

SPECIALTY SECTION
This article was submitted to
Computational Psychiatry,
a section of the journal
Frontiers in Psychiatry

RECEIVED 08 November 2022
ACCEPTED 16 January 2023
PUBLISHED 07 February 2023

CITATION
Castro Martínez JC and Santamaría-García H (2023) Understanding mental health through computers: An introduction to computational psychiatry.
Front. Psychiatry 14:1092471.
doi: 10.3389/fpsyt.2023.1092471

Computational psychiatry recently established itself as a new tool in the study of mental disorders and problems. Integration of different levels of analysis is creating computational phenotypes with clinical and research values, and constructing a way to arrive at precision psychiatry are part of this new branch. It conceptualizes the brain as a computational organ that receives from the environment parameters to respond to challenges through calculations and algorithms in continuous feedback and feedforward loops with a permanent degree of uncertainty. Through this conception, one can seize an understanding of the cerebral and mental processes in the form of theories or hypotheses based on data. Using these approximations, a better understanding of the disorder and its different determinant factors facilitates the diagnostics and treatment by having an individual, ecologic, and holistic approach. It is a tool that can be used to homologate and integrate multiple sources of information given by several theoretical models. In conclusion, it helps psychiatry achieve precision and reproducibility, which can help the mental health field achieve significant advancement. This article is a narrative review of the basis of the functioning of computational psychiatry with a critical analysis of its concepts.

KEYWORDS

computational psychiatry, computational phenotype, precision psychiatry, translational psychiatry, computational modeling

Introduction

The brain has been conceptualized as a computer performing continuous calculations about itself and its environment. Moreover, according to the theory of systems and Bayesian approaches, the brain is conceived as a complex, non-linear computational device (1, 2). The mentioned approaches could benefit a further comprehension of multiple levels of analyses that subsume mental health and psychiatric diseases.

In the psychiatry field, various attempts have been made to understand mental health and disease fundamentals. However, those intents have generated different explanations within multiple theoretical models, which are often disconnected and lack of complex understanding of mental health and psychopathology integrating many levels and systems. Thus, a point has been reached where a paradigm shift is needed. Dimensional and transdiagnostic levels of understanding are required to better comprehend. Some of the possible answers have chosen the use of mathematical principles to reach a multilevel analysis and generate hypotheses that can be validated. Such an approach provides the possibility of achieving a unifying theory, increasing accuracy, and reproducing what was found previously by other authors. In this context, computational psychiatry is a tool for precisely this purpose. It should be clarified that this probabilistic view of the brain is open to controversies (3).

Psychiatry has always encountered multiple controversies during its history. These, in turn, have generated multiple internal and external crises that have questioned its validity as a science and its management of mental illness (4–6). These criticisms have focused primarily on the validity of their concepts and constructs (7), their diagnostic capacity (8), the reliability between different observers, and the lack of biomarkers to determine the diagnoses, treatments, and prognosis of the condition (9, 10). Additionally, they have focused on the variability of the course of different disorders, typically heterogeneous in their presentation (11).

Psychiatry has used various approaches to overcome these criticisms, such as nosological formulations. This strategy attempted to elucidate their biological basis (12) by achieving greater reliability in the diagnosis. Such systems have generated multiple syndromes with significant heterogeneity in their course, clinical manifestations, prognosis, and response to treatment but grouped under the same diagnostic category (13). This has raised the possibility that they also have a different neurobiological basis. It has also shown the limits of these tools. However, a myriad of empirical data has been obtained through such systems. Although, this data suffer from poor integration of cellular, synaptic, neuronal circuitry, and complex behavioral responses (14, 15). For example, there are strong interactions between genes and environment at the genetic level, but no clear paths to how these develop into a specific phenotype. This also occurs in neuroimaging, where only indirect measurements of the behavioral variables observed in clinical practice have been achieved (15). In conclusion, a cohesive model capable of taking data from different sources and giving adequate weight to each source of information has yet to be reached.

Computational modeling of behavior was elaborated by specializations of the neurosciences, which preceded computational psychiatry. One of the first to do it was computational neuroscience. It is responsible for studying the brain at a theoretical level, determining the principles and mechanisms that guide the development, organization, and process of information (16). This is achieved using computational models (descriptions and explanations of processes) that occur at different spatial and time scales and with non-linear interactions. More specifically, computational neuroscience makes hypotheses about the processes that operate in the brain at different analysis levels and unites them to corroborate them. The goal is to understand the functionality of complex systems such as the brain, formulating quantitative hypotheses (17). In this context, computational models give a practical tool to address specific brain characteristics, such as its emerging functions. Depending on the question type one wants to answer; one will opt for a particular abstraction level to form a model. Here, the three levels of analysis proposed by Marr and Poggio are relevant (18). These are the computational level (the "why"), which is the most abstract and deals with logical-mathematical reasoning; the algorithmic level (the "what"), which evaluates the rules of the process; and, finally, the level of implementation (the "how") (19). For this author, brain research was conceived as a problem of information processing (17).

Based on this, computational psychiatry has appeared as a way to achieve this integration. Computational psychiatry uses formal models of brain function to characterize the mechanisms of different psychopathological manifestations by describing them in computational or mathematical terms (20). This facilitates the study and articulation of these data by incorporating knowledge from other sciences such as cognitive science, computational neurosciences, and "machine learning" (20–24), trying to translate knowledge between different levels of analysis. This review aims to give a comprehensive view of the foundations of computational psychiatry, highlighting its interactions with different approaches like biophysics and evolutionary psychiatry to arrive at precision psychiatry.

This field has become an essential tool for finding novel solutions, encompassing both the context and the individual. In addition to providing investigative and practical means to arrive at response to specific needs in these contexts in a cost-effective way. Nevertheless, it is necessary to understand its foundations and how it applies to research and clinical purposes.

In this scoping review, we first describe the importance of computation modeling in psychiatry to face limitations from a system theory perspective. Then, we explain how computational models are built, giving particular emphasis on their underlying concepts. Moreover, we comprehensively explain the statistics surrounding the computational models and their applications at different levels of plausible explanations in psychiatric scenarios. Finally, we reflect on model validation and the potential limitations of computational psychiatry.

Methods

A narrative review of the literature was conducted, focusing on computational psychiatry's fundamental concepts and applications. To this end, we searched PubMed, MEDLINE, EMBASE, and EBSCOHost for both narrative and systematic reviews of computational psychiatry using the terms "computational psychiatry," "biophysical psychiatry," "computational modeling," "digital phenotyping," "precision psychiatry," and "computational neuroscience." After this search, essential studies were also reviewed within the articles' references. Articles written in English and Spanish were selected. Articles based on their publication date were not excluded. The last search was conducted on 30 December 2022.

Need for comprehensive models of mental illness

A model is a heuristic way of understanding complex interactions and their relationships by employing a simple rule (25). In the case of mental health, modeling used have degenerated into a diversity of disjointed data from various theories. Additionally, modeling through the relation between brain activity and psychiatric phenotypes has many pitfalls because they are focused and predict complex profiles rather than unitary cognitive processes (7, 25).

Most explanatory models used in psychiatry and psychology focus on narrative methods, with the problem of approaching human behavior from only external behaviors or epiphenomena conducts without ever finding a clear biological causal or mechanistic basis (7, 25). This approach leads to difficulties in determining clear biological and clinical processes (25) with implicit categorical errors (26). One example is the measures based on self-reports with poorly defined variables and poorly elucidated pathophysiological mechanisms (14). On the other hand, by not knowing the mechanistic or robust theoretical approaches to study mental disorders, some studies are initiated to look for relationships with multiple variables in so-called "fishing expeditions." In turn, this can generate associations that do not reflect the actual phenomenon (27).

Moreover, theoretical models have often been chosen for data-driven approaches. However, this theoretical approach can exhibit challenges when studying human behavior as they assume *a priori* hierarchies assessing predictors of an outcome and can be restricted to a partial understanding of a complex model (25). Finally, one can obtain data replicated by others, assuming a certain degree of validity, which could ultimately be wrong. In fact, replicability issues represent one of the biggest current challenges in psychological studies (25).

Computational models applied in neuroscience and psychiatry

Neuroscience and psychiatry lack methods for constructing, assessing, and validating theoretical models, which should be more extensive than describing relationships between different variables (25). Against this issue, computational modeling of dynamic systems becomes vital as it allows the generation of data-driven validations of conceptual reference frameworks and biological measures, avoiding issues due to spurious statistical associations and biases in building models (25, 28).

Thus, computational models help simultaneously manage massive information sources and articulate biological, psychological, and contextual models for understanding human behavior. Different approaches as machine learning, deep learning, and explanatory modeling (13, 29), help to process information building in models only determined by data avoiding theoretical and restriction biases.

An important method computational psychiatry uses is differential equations, which express neurobiological systems' functioning more closely. They represent changes as a function of time codified by the interactions with other non-linear variables (25). Consequently, they can join several equations that mathematically specify relationships between symptoms, environmental factors, and neurobiological substrates (28, 30, 31). This exemplifies the possibility of reconciling different perspectives and empirical data, providing cohesive, stratified models for understanding complex phenomena.

A big group of mental disorders computational models focuses on altered learning and decision-making processes as the central components (15), highlighting the relevance of information integration processing. These learning models have been used in computational cognitive neuroscience, using tools like machine learning to model a specific phenomenon. These cases are usually divided into supervised, reinforcing, and unsupervised. Within these models, it is assumed that the objective of learning is to form storages of representations to be remembered and guide behavior, although the mechanisms to perform it may differ according to the model (29). In the case of supervised learning, specific feedback is received after the experience.

In contrast, in reinforcement, this feedback is not explicit and can be delayed and influenced by multiple factors. In addition, it can be done in the form of punishments or reinforcements, which may not be directly associated with the behavior. Finally, in unsupervised learning, the subject is the one who must make sense of the experience without any feedback.

Another use of the computational approach in psychiatry is modeling a specific phenomenon. To achieve this, computational models offer a tool to facilitate it *via* the generation of self-generated models that synthesize data through the sampling of inputs and achieve an approximation of specific outcomes, thus integrating Bayesian probability (15, 32–34). Ultimately, they enable us to make a probability distribution and hierarchization of the best predictors of a neural, cognitive, or behavioral state among massive interactions from different sources of variables (27). In psychiatry, mentioned models would help us elucidate and better understand psychopathological phenomena by relating them to neuronal processes and their normal function (35).

This step can be performed at different levels of explanation, such as at the molecular level, neural networks, cognitive processes, or mental symptoms. Computational models now allow us to predict symptoms and clinical presentation of neuropsychiatric disorders by studying brain volume information and functional connectivity networks *via* data-driven methods (machine learning procedures, support vector machine methods, or deep learning approaches) (32–34). Moreover, these approaches have been relevant for biophysical psychiatry (14), where psychiatric phenomena of interest, such as psychopathology, relate to alterations in the biophysical properties of, for example, the membranes of neurons, as seen in other reviews (14).

Moreover, computational methods have helped to understand the impact of different sources of information in neurodegenerative disorders (36) or elucidating individual and contextual factors determining complex behaviors such as violence (37).

Statistical foundations

Computational theories of the mind are based on probabilistic perspectives. The brain processes are considered mimicking computational functions of the system to infer the state of its environment and decide which course of action to follow (35, 38). The inputs will never be completely reliable, so there will always be uncertainty that has to be considered when performing any task. Therefore, Bayes' theorem (the combination of the initial expectation of the state of the environment and the probability of the input determining a modified estimate of the state of the environment) is used to describe these processes (15, 35). Describing in such a way brain processes can be translated to computational psychiatry approach. This contrasts with the statistical approach used in psychiatry, which asks about the probability that the data have resulted from the null hypothesis (25) and corresponds more to discriminative models (35). By contrast, in the computational psychiatry perspective it is possible to assess different layers of biological, psychological, and social-contextual information and use algorithmic approaches to assess multiple interactions between layers, modeling data and testing those models with complex validation processes of findings.

Different interactions can be found when computational approaches are studied in neuroscience and psychiatry. First, some computational models in neuroscience accept the metaphor of the brain as a computer (20, 39). Mainly, the models who accept the metaphor of the brain as a computer describe brain biological processes as part of a computer that primarily formulates predictions on future states based on massive integration of past interoceptive and exteroceptive information. In this perspective, the brain is an entity that constantly builds and updates a model of reality through sensory inputs named generative (40). The optimization of this generative model must lead to the minimization of free energy (the energy used by the brain) (2). Moreover, the brain tends to formulate different predictions and minimize errors to curtail energy expenditure, which can be done in two ways: either by adjusting the cognitive scheme of the world or by changing the pattern of action (15). The latter is essential since it can explain psychopathology, including functional neurological symptoms (41).

Computational psychiatry is aligned with previously mentioned notions, as it integrates different levels of information to formulate appropriate models to describe and understand mechanistically healthy and pathological behaviors. Moreover, computational psychiatry performs predictions of potential states and biomarkers and runs test and retest validations assuming complex heuristics to predict psychiatric phenotypes (18, 35, 42). The goal is to generate accurate and robust predictions with the minimization of the workload to reach meaningful outcomes.

According to Breiman (43), two statistical models are used in the mentioned approaches. The first is algorithmic or "data-driven" models, aiming to predict results by having a specific group of data (inputs) following complex statistical procedures leading to massive interactions between variables. The second model is described as "theory-based" modeling, where a pattern of outputs and initial data is used to determine how the process is performed to generate this data (35, 44).

Both types of statistical analysis proposed by Breiman share statistical tools and can be associated with concepts from the learning field through reinforcements. Thus, it offers different visualizations as to how to conceptualize the computations or calculations made by the brain (15). Also, it can give way to the use of methods like machine learning in computational psychiatry (21). Learning is a complex process since there is always uncertainty. A specific behavior is selected according to the reinforcers and punishments values during reinforcement learning to maximize a particular outcome. All this is based on the prediction error measured utilizing the learning rate. The impact of this error depends on its accuracy (inverse uncertainty) (45).

There are two different ways in which experience is used to estimate and predict future rewards and punishments. The first is a model-based cognition, also called goal-directed, where experience is compiled into a generative world model. This involves the inference of future possibilities, generating an enormous computational cost. This contrasts with model-free cognition, where no information about the change suffered is stored but only encodes how much reinforcement is obtained when the subject is in a state or performs a specific action. In the latter, computational costs are decreased, but at the cost that the system becomes slow and inflexible, with no possibility of responding to changes in the environment (46).

To abridge previous gaps, the predictive coding model is important, where a unit at a specific hierarchical level sends messages to one or more units of lower levels that predict its activity (47, 48). The discrepancies generated between these predictions and the actual input are then passed to higher levels of the hierarchy as prediction errors. They are then reviewed to refine the prediction (35). The uncertainty (inverse precision) of each level determines the rate of learning at each level, determining the size of adjustments that must be made to explain the data that has been sensed. This approach is closer to the representation of the nervous system, a dynamic and hierarchical system. However, this hierarchical model has come into question with models such as the heterarchical model (49, 50), where the components of a system do not have a specific order. However, they can have different connections depending on the function and the context that is being analyzed.

Statistical models based on data could give important tools for clinical practice. One is SpeechGraph, a computational tool that can quantitatively assess a patient's discourse structure through graph theory (51, 52). This tool does not take the process of speech formation (the syntax). However, it is possible to calculate the attributes of the graph created from the discourse and, through these, can differentiate a control from an affective and non-affective psychosis (52–54). It can also determine the differences in the development of the discourse longitudinally of children with psychosis and controls (55) and in cases of dementia, these can be correlated with other cognitive deficits (56). The importance of these approaches has been taking force with Natural Language Processing (NLP) associated with machine learning paradigms (57, 58). These models can evaluate specific parts of a complex neurocognitive process like language and then aid in comprehending the underlying pathological mechanisms.

In the statistical models used in theory-based models, the parameters are surrogate variables of neural computations (processes). In this case, the parameters do adapt to neurological or behavioral data. Therefore, this model can be used to elucidate possible dysfunctions underlying multiple mental disorders (59), such as the search for pathophysiological processes underlying transdiagnostic alterations. This theory-based approach can also account for neurocognitive approaches like the Bayesian active inference model of discourse. The person speaks, and this person monitors internal and external signals in the search for errors (60) and explains the way social cognition alterations could disrupt language emission or reception.

To this end, several conditions must first be secured. The first thing is that the model must be able to predict multiple experimental data. To determine this, the effects of the parameters on the model's predictions must be independent, and there must be sufficient data. These conditions are mainly used to compare different models to determine which fits best to a particular phenomenon. One of the ways to do this is to simulate data in each model one has and determine the ability of each model to generate the "real" data of the variable being studied. Remember that the empirical and predicted data will not coincide perfectly (59). Then, these parameters may be used as computational markers of psychiatric illnesses. These markers associate psychiatric dysfunctions with failures in neuronal computations (predictive coding, divisive normalization, and contextual modulation). With these markers, what one is trying to do is (59): (i) Distinguish between diagnoses with similar symptom profiles: spectrum problem or symptom overlap. (ii) To characterize heterogeneity within diagnostic categories concerning alterations in computational mechanisms. (iii) Predict relapses or responses to treatments.

Although brain processes result in great complexity, they present a hierarchical organization that allows them to be broken down into more basic operations and easily understood. In the same way, phenomena studied by psychiatrists can be simplified and organized in hierarchical models through factor analysis (61) or network theory (62).

Levels of analysis

Overall, this data processing method seeks to integrate neurological, psychological, and social reference frameworks. Indeed, it searches for a way of making bridges between different levels of this hierarchical organization of the brain, which has to consider its surroundings as described by the concept of the *phantastic organ* (21).

It is then possible to make models that describe the molecular basis of the individual neuron, describing its electrical properties and the generation of the action potential. This is done by employing a set of equations that describe its properties (14). To achieve this level of characterization, previous studies of significant impact on this understanding are taken as the description of the signal propagation by neurites (63) and others (64).

Also, within the first level of analysis, one can opt for the genomic analysis and description of the studied phenomena. Genomic approaches attempt to determine the biological relevance of genetic variants and predict their influence on the phenotype (65). This review shows that computational models lend themselves precisely to validating and confirming biological relevance. Currently, the discovery of possible risk variants using GWAS (66–71) is much faster than their validation. Nonetheless, computational approaches have been developed for the prioritization of disease-gene candidates (72). This advancement has enabled researchers to elucidate co-expression patterns through network analysis (73).

Before continuing, some clarifications must be made regarding the conceptions of circuits at the neurosciences and the clinical level. The term circuit in neuroscience refers to microcircuits where biophysical processes modulate a response. Meanwhile, in clinical neurosciences, these are dynamic systems defined by control systems (25). This latter definition describes better the networks which are studied in psychiatry.

However, to discover circuits associated with a specific phenomenon, a hybrid approach must be used where the discovery of a circuit is based on observing the dynamics of its outcomes. This is then put into a differential equation that describes the system's mechanistic structure. With this, a differential equation that describes the system's response can be generated starting from the inputs and outputs. This is called the "transfer function." At this time, machine learning can be used to discover plausible or related biological circuits. These circuits can be added according to their interactions generating complex systems (25). This kind of approach enabled researchers to develop theories of the function and associations of specific brain regions like the hippocampus (74) and, with this information, able to put forward hypotheses of different pathologies (75).

At the level of circuits, an attempt is made to elucidate the intrinsic neural activity evoked through different brain systems. These models incorporate the properties of neurons and synaptic connectivity. However, they are limited by the strategies used to acquire information from these networks, based on imaging studies. So, these models describe the brain as a network of interconnected nodes (76). To achieve this, there is a necessity to describe the structural connectivity matrix together with an equation that determines the neural dynamics of each node. Direct connections and the background activity of the area will influence these. Many of these aspects require biophysical knowledge at the molecular and cellular levels to achieve a more accurate approach to empirical neural dynamics. Furthermore, they could integrate with data and knowledge taken from the connectomics fields. In doing so, these approaches are helpful in investigating alterations in the brain connections in specific diseases (77) or arrive at transdiagnostic alterations (78).

As for the psychopathological level, examples are scarce. However, computational psychiatry can also be used to reach its understanding and even form practical applications based on psychopathological alterations of the computational level of information processing (79), such as salience processes. This opens the possibility of evaluating neurocognitive domains to evaluate a patient, which is currently underused for patients with psychiatric ailments. Within this part of the diagnostic and therapeutic process in psychiatry, various problems previously highlighted in terms of the validity of the psychopathological evaluation and nosological classification can be tackled (4–6, 80–84). Within this panorama, psychopathology can be considered a complex system (85), where alterations in its balance generate a search for homeostasis through an orientation toward the environment and a manipulation of its parts, reaching emerging qualities, which can be expressed as symptoms during a mental examination. Because of this, computational processes are privileged to achieve new perspectives that allow the clinician or researcher to overcome these obstacles.

Moreover, there is a possibility of considering constructs that may not be psychopathological but do contribute to suffering, such as domestic or gender violence. Nonetheless, different models have been proposed to tackle this problem, like the Hierarchical Taxonomy of Psychopathology (HiTOP) (86) and the network theory (81). This is how different diagnostic approaches using computational methods have been proposed (71).

However, to achieve this, different levels of analysis contribute differently to a specific phenomenon. Nonetheless, their integration is difficult to achieve, as well as the identification of a level more essential to the phenomenon studied. So, depending on the question to be answered, specific methods must be used to address it. Consequently, certain analysis levels will also be used preferentially. The problem lies in recognizing which level permits having a bigger and better picture of the studied event, weighing each component differentially. In other words, according to the question to investigate, certain elements of the phenomenon will be more important (essential) than others (87). In such a way, certain levels of analysis will carry more information within this question.

Finally, this must also be complemented with a longitudinal perspective (88–91), in which importance is given to how these processes will shape neurodevelopment (92, 93), where both normal and abnormal trajectories of such development must be studied for the possible determination of useful biomarkers or the understanding of the interactions that are at play and that can be associated with both normal and abnormal development. All this is associated with perspectives promoted by the RDoC initiative.

Considering the above, there are three types of perspectives to approach the description of dynamic systems, such as mental processes (25):

1. "Bottom-up" biophysical approaches: begin in individual neuronal functioning and are extrapolated to other levels of

hierarchical organization, such as networks. In this case, the equations represent the properties of neurons, synapses, or ion channels. These observations can then be transferred to the functioning of neural circuits. However, the subsequent step between circuits and behaviors is much more complicated to surpass. There is an underlying problem: the whole can be greater than the sum of its parts. So, understanding the basic processes does not always arrive at a corresponding process in the higher levels of the hierarchy.

2. "Top-down" approaches, where one starts with an emerging phenomenon and tries to infer the set of neural mechanisms on which they are based. In this part, connectionist models used in cognitive and psychiatric neuroscience become essential. These models study neural systems that are involved in various cognitive processes. They attempt to arrive at the functioning of neural networks on a large scale and thus be able to achieve behavioral predictions (12). Models incorporating more than neurobiological systems are included in this part of modeling, such as social interactions or cultural influences.
3. Theoretical-informative approaches, where structural strategies are investigated where the brain can optimize the efficiency of information propagation based on graphs or network theory considerations.

Now, this complements modeling levels suggested by computational neuroscience (35), described in the introduction.

Model's validity

Validation requires integration of multiple sets of data involving different biological, psychological, social, and contextual levels of analysis. Computational modeling in psychiatry maximizes the amount of information predicted using data-driven hypotheses and testing processes *a priori* assumptions using a small pool of data (25). Another advantage of computational psychiatry is the capacity to generate, test, and validate available models or those generated in the research process (25).

Modeling validation can be tested using statistical parameters, including accuracy, sensitivity, specificity, and power measurements. The accuracy of a model is critical to take into account as it allows a certain degree of confidence in the results obtained. This is determined by the degree of error between the prediction it makes and the empirical data obtained. Meanwhile, the model's power is evaluated by the diversity of inputs (different perspectives) and the time during which the predictions are valid (25).

Thinking of the brain as a machine that solves inferential problems can be an excellent way to generate testable computational hypotheses about psychiatric disorders (35) or even mental issues. Moreover, this is especially important because each measurement can have multiple explanations (multi-causality). The problem lies in finding which description is the one that best fits the data taken and enables a better prediction of future problems. For this step, the researcher can determine the model parameters that maximize the likelihood of the data given the model in a process known as model fitting. This likelihood is then used to calculate a quality-of-fit criterion (94). There will also be a degree of uncertainty, and there will always be room to improve the models. Then, a balance must be made between the complexity of the model and the model's accuracy. A highly complex model leads to greater difficulty in achieving the understanding one wants to have of the phenomenon studied (95). But mental processes are highly complex, and some complexity of the model is inescapable. Alternatively, simple models can lead to poor prediction, in other words, lower accuracy and thus low usefulness. In addition to this, it must be considered that psychiatric disorders are characterized by their heterogeneity, so there may be several mechanisms at play in the same patient despite having the same phenomenological or nosological representation, which must be considered within the validation process (15).

As this computational modeling field grows, there is also the need to be able to compare different models. One such way to do so is using Occam's law (94). Similarly, as the free energy principle governs the brain, one could select a model according to its predictive performance (its ability to predict observed data). Nonetheless, this approach is not enough for selecting theories. In this scenario, the model's generative performance becomes a better way of selecting the model by falsifying it (94). The latter requires the simulation of candidate models in a denominated model recovery process. These two selection models are complementary an allow researchers to reach the most accurate modeling to explain a dataset (94).

Finally, the greater complexity and computing power put forward another issue: reproducibility. For an article to be reproducible needs that researchers share its data and coding, and in executing the code with the data given, one arrives at the same results. The ability to analyze more complex interactions between non-linear factors and their dynamic interplay could bring the researcher closer to data with a low signal-to-noise ratio, with a possibility of identifying false associations (96). It is essential to point out that this is not unique to computational approaches. However, it is partly facilitated by multidimensional datasets which go through rapid, flexible, and automated analysis (97, 98), as in Big data approaches. To tackle this problem, sophisticated analyses are required. However, there is a lack of infrastructure and knowledge to support this task.

Nonetheless, initiatives have taken place to tackle these limitations, and various articles have been written to describe steps to take to achieve the goal of more reproducible research, like improving methodological knowledge and independent methodological support with the encouragement of collaboration initiatives and open science (97) and to develop a way of accountability (99). It is also important to bear in mind the bias-variance trade-off (100). There is a conflict between bias error and variance error which must be minimized while constructing a computational model. A bias error generates when the model is not capturing relevant associations, while a variance error occurs when the model is overfitting.

Precision psychiatry

When talking about precision psychiatry, we seek to achieve a computational phenotype. This means achieving a model that best suits the empirical data of the subject or phenomenon. This allows for generating inferences at the individual level about the underlying computational mechanisms that govern what is observed in the patient, thus overcoming the opposition between the dimensional and categorical perspectives (12). This is of utmost importance in ethnopsychiatry since it allows to the generation of specific modeling of behavioral alterations, which can be outside the nosological categories. Equally important, they acknowledge the impact of specific environments in a person's life. In doing so, a better understanding of the person and their contextis reached; and one is capable of offering the best possible therapeutic approach (individualized and person-centered).

However, the traditional form of research in psychiatry has allowed predictions of the average functioning and mechanisms of pathophysiology to be achieved in a defined group of patients, such as that presented in nosological systems. Nevertheless, the problem of proposing differential diagnoses arises. When a differential or comorbid diagnosis is suggested, the clinician must determine from the findings in the patient what is the specific pathophysiological mechanism or of more significant predominance in the individual (15), which is currently impossible. This would determine the best therapeutic intervention for the patient and their prognosis (101). This possibility of differentiation between diagnoses and spectra within the same diagnosis has been made possible through "generative embedding" (102), although only in research.

Another problem is to consider phenomena outside nosological systems, which also have a high impact on society. A clear example in the Colombian case is that of violence, from which multiple phenomena and complex social processes have been generated that have contributed to the mental health of a population (103). Still, they should be given more importance in the research on mental health, especially from the medical perspective (104).

Clinical applications

Transferring all the previously described concepts to the clinical and practical field has been costly and time-consuming (105). It is one of the most critical efforts to test the usefulness of these approaches (87). This has multiple reasons, which could be summarized as that mental health depends on normal brain function and how it is related to modification and is influenced by the individual's context. It is a form of circular causality. These models or tools must describe dynamic, hierarchical, and non-linear systems. This means that it is challenging to have a clear and concise understanding and comprehend these phenomena or disorders. However, approaches are trying to address this problem by creating a bridge between neuroscience and computational psychiatry with cognitive neuroscience. It is essential to highlight that computational psychiatry can be a valuable tool in searching for these basic computations and how they modulate and emerge innovative functions from an evolutionary perspective (38).

Currently, psychiatry is primarily based on nosology contingent on classification systems such as the DSM or the ICD. However, this approach can be complemented by a dimensional vision, where they are added to the psychopathological manifestations and dimensions given a value within a continuum in models like HiTOP (86). However, this value can be non-linear or interact or correlate with other dimensions by modifying the syndrome and making it extremely difficult to quantify the weight of a specific factor.

A clear example is the determination of suicide risk (106, 107). The risk factors are determined through previous studies, but the quantification of these is carried out at the level of the clinician's judgment, and the scales have poor operational characteristics (106–108). In addition, all this is done from population data without considering the differential influence of these factors on the individual. Machine learning has been used to predict suicide attempts and deaths from clinical records (102). For this reason, a way is required in to integrate dimensional and categorial visions, which often escapes the possibility of the clinician within their daily practice (44). The difficulties in diagnosing, prognosis, and treatment of this type of patient are highlighted. To have a complete picture of these applications, the reader can refer to the review made by Huys et al. (44).

These first approaches are still only applicable to research, but they give glimpses of the utilities of this tool. On the other hand, one can have clear examples where the first steps have already been taken to achieve a translation of this knowledge. Some of these examples are available below.

Data-driven approaches

1. *Diagnostic classification:* In this aspect, elements of "machine learning" can be used. With this, neuroimaging data can be analyzed by distinguishing clusters of specific symptoms with specific neurobiological substrates, as seen by Costafreda et al. (109) or Mota et al. (54, 110). However, problems such as determining comorbidity as completely different disorders continue without the possibility that they have defined diagnostic limits (111). Because of this, the usefulness of these tools requires testing their properties in ambiguous cases, where there are more significant difficulties in differentiating.
2. *Prediction of clinical status*: This type of application focuses on identifying markers to determine the stage where a particular patient is to describe prognostic or treatment features. This has been used in early psychosis to predict social outcomes in a high-clinical risk sample (112). In other examples, NLP can be applied to clinical records like psychotherapy notes to enhance prediction models for different clinically relevant outcomes like suicide risk (113).
3. *Prediction of treatment response:* This aspect corresponds to the need to improve the prognosis and the ability to identify the best therapeutic alternative with an individualized approach. In the specific case of depressive disorder, where it is evident that only two-thirds of patients have a response after multiple pharmacological attempts (114–116), identifying the characteristics that could collaborate in the treatment choice is required. It may be that the cases referred to as resistant are not but require differential therapeutic responses. However, it has been attempted to achieve different ways to characterize and predict treatment responses, such as quantitative electroencephalogram markers (qEEG) (117, 118) which were validated by other studies (119). In addition, methods based on neuroimaging results have also been used (120), which be associated with computational approaches for pattern classification. All these approaches have been shown in their early experiences to improve responsiveness.
4. *Choice of treatment:* As mentioned in previous section, not all patients respond in the same way to treatments, even if they are first line. But as made explicit above, there are no variables or individual characteristics of the patient to determine it, even though multiple pharmacogenetic studies have been done in some specific situations. At this point, numerous binary classifications can be used simultaneously to achieve this task. However, to be feasible, a specific group of paraclinical must be

used (121, 122). It can be used, for example, in electroconvulsive therapy, where simulations of electric fields can be integrated with the current knowledge of neurocircuitry to individualized electrode configurations (123, 124) and in the Deep Brain Stimulation (DBS) field (125).

5. *Clustering of clinically relevant data:* In this approach, unsupervised methods are used to cluster together characteristics of the sample giving rise to dimensional factors that can inform the patient's clinical status. An advantage of this approach is that it facilitates the interpretability of the results (100). This approach has been used to identify brain fingerprints in different disorders from neuroimaging data (126, 127).

In these different applications, the researcher can take various sources of information to give a more accurate picture of the patient (128). This complementarity exemplifies the possibility of the constructing of mechanism-driven knowledge from data-driven approaches (100). These applications could then be articulated with network theory to understand mental disorders revised elsewhere (62, 129).

Theory-driven approaches

These are initially "fed" by multiple data found at various levels of research, exploring the relationships between them. At the level of psychopathology based on Bayesian theory, the psychopathologic symptoms can be structured in three different ways: solving an inappropriate problem correctly, solving a suitable problem incorrectly, or solving a relevant problem correctly but in the wrong context (130). Moreover, from this conception, an analysis and a possible union of knowledge of brain structure and functioning can be generated together with behavioral variables seen in clinical practice.

1. *The course of the disorder:* In this section, Goldbeter's article can be an example (131). The author gives a model of mutual inhibition between two processes (depression and mania) to explain the cyclicity seen in the disorder. In this example, the model does not contemplate neurobiological processes at neurocircuits, synaptic, neuronal, or biophysical levels. Still, it achieves a conceptualization of a phenomenon of extreme importance, such as the cyclicity in bipolar disorder.
2. *Predicting risk of recurrence:* There are other examples where the researcher could take a specific marker like effort and reward tasks to determine the clinical status of a specific disorder. And later, decide on the treatment according to the information this marker gives the practitioner, like the risk of recurrence (132).
3. *Neurocognitive functions:* The models can be used to describe the function of neurocognitive functions and domains, like working memory (133). And it enables researchers to put forward theories and models of pathological alterations of these processes (75). These models can also be used directly in conceptualizing a disorder like obsessive-compulsive disorder (OCD) and linking it to neurodevelopmental processes (134).
4. *Pathophysiological processes*: This type of model can give insightful perspectives that integrate different levels of analysis giving rise to a comprehensive and integrative knowledge of the disease processes. There are multiple examples across multiple disorders like schizophrenia (135–137). This, in turn, could give information about possible therapeutic targets. Researchers could also create models for explaining and understanding mechanisms associated with the therapeutic response, like neuromodulation strategies such as ECT (138).

In addition, an integration of these two approaches can also be achieved. This is because theoretical models must be fed from previously collected data to construct a good model. But also, a mechanistic model can generate available data for constructing pragmatic tools that can be used in clinical practice. To show how this applies to a specific pathology (schizophrenia), refer to the article by Valton et al. (139). In **Table 1**, there is a list of the examples used throughout this review with a description of the approach used and implications and contributions for the field.

Finally, it is essential to highlight that these applications go beyond the nosology provided by the DSM and allows the visualization of phenomena that can impact the course and prognosis of these disorders or the mental health of individuals in general. An example of this is creativity, which can be understood as the ability to create unique products such as artists (Creativity with a capital C); or as a cognitive function that helps the individual adapt to his environment and give answers to his environment (creativity with c) (140). The latter, in turn, depends on divergent and convergent thinking (141). In the review carried out by Mekern et al. (142), it can be evidenced how the same phenomenon can be studied from different levels and segmented into other processes even going so far as to predict or determine how these processes would be affected by specific alterations or disruptions. With the help of computational modeling, it improves its understanding.

RDoC: Possible response to the constraints of nosological systems

The nosological systems encountered in the clinical and research practice delineate highly heterogeneous phenotypes that lack reliability and validity, which has restrained advancements in the field as the computational tools rely on the input one puts in them (82, 143). In this way, if one takes invalid or erroneous input to a model, which can be valid, the data that results from this process is also invalid and could deviate the researcher to a categorical error. The necessity for a system of categorizing these problems and disorders in a way that conceptualizes them as a mixture of interacting and dimensionally varying processes is at the front and center of the problem (87). This, in turn, could give us a way of representing these problems in a more ecologically valid way.

Different approaches have been made by researchers in order to arrive at solutions to these limitations. One of them is RDoC. The *Research Domain Criteria Project* was initiated by the NIH (National Institute of Health) to address the different problems that research has encountered in mental health, specifically mental health disorders (144–146). This project was conceptualized as a research framework, so it has no applications in clinical nosology, nor does it pretend to be a replacement for it. Although, one of the potential impacts is to achieve a classification system with a more significant neurobiological basis without leaving a biological reductionist vision of these

TABLE 1 Clinical applications of computational approaches.

Title	References	Study aim	Data analyzed	Computational approach	Conclusions/Implications
First symptoms and neurocognitive correlates of behavioral variant frontotemporal dementia.	Santamaría-García et al. (32)	Analyze neurocognitive correlates of patients with bvFTD who debuted with apathy or disinhibition.	Data from a group of patients and controls involving neuropsychological, clinical, and neuroanatomical data.	Data-driven approach using machine learning associated with a multivariate analyzes.	This study gives an example of the possibility of integrating different levels of analysis of data with a longitudinal perspective. The latter is achieved by the longitudinal approach to the study and the description of correlations of first symptoms and their evolution. This study assessed multiple levels of analyses by implementing support vector machine approaches.
Robust automated computational approach for classifying frontotemporal neurodegeneration: multimodal/multicenter neuroimaging.	Donnelly-Kehoe et al. (33)	Determine if by using atrophy and resting-state functional connectivity one could differentiate between patients with bvFTD and controls.	Datasets from participants in different regions of the world.	Automatic, cross-center, multimodal data-driven computational approach using machine learning.	The multimodal approach explored in this study enhances the system's performance in a multicenter protocol. This underscores the possibility of clinical applications in real-world conditions. This study implemented different machine learning models to abridge different levels of neurocognitive and clinical information in dementia.
At the heart of neurological dimensionality: cross-nosological and multimodal cardiac interoceptive deficits.	Abrevaya et al. (34)	Examine the impact of neural relative to autonomic disturbances of cardiac interoception across neurological conditions.	Data from 149 participants divided between two pathological groups (neurological and cardiac) and controls.	Data-driven approach to evaluate the relevance of the cardiac interoceptive dimensions in the discrimination of neurological and cardiac pathologies. A classification pipeline was used with the input from behavioral dimension and different levels of analysis.	This study demonstrates the possibility of computational models to integrate different systems (cardiac and neurologic) to find relevant variables for the discrimination of disorders. This study reached to mentioned conclusions by implementing different automatized analyses including support vector machines and machine learning procedures.
Thought disorder measured as random speech structure classifies negative symptoms and schizophrenia diagnosis 6 months in advance.	Mota et al. (54)	Determine if early markers of speech disorganization during recent-onset psychosis measured using SpeechGraph could correctly classify the severity of negative symptoms as well as the schizophrenia diagnosis.	Graph measures of different memory reports.	Data-driven software to measure graph attributes of connected speech.	This study has a different approach to the use and application of computational models. It takes a software made through a data-driven approach to arrive at quantitative measurements of formal thought disorder. This could in turn help to delimitate better these alterations to make a more precise diagnosis. There are other applications of this software (53, 55, 56, 110, 155, 156).
A computational framework for the prioritization of disease-gene candidates.	Browne et al. (72)	Evaluate the performance of a method for gene prioritization applied to Alzheimer's disease.	Gene Expression Omnibus (GEO) database.	Model-based approach based on network theory for the creation of Protein–Protein Interaction Networks (PPIN). Integration of multiple datasets for the construction of PPIN.	A framework that integrates diverse heterogeneous data including gene expression and network topological features to prioritize and analyze disease-gene candidates applied to AD as a Case Study. Demonstration that the integration of PPINs along with disease datasets and contextual information is an important tool in unraveling the molecular basis of diseases.

(Continued)

TABLE 1 (Continued)

Title	References	Study aim	Data analyzed	Computational approach	Conclusions/Implications
Integrated co-expression network analysis uncovers novel tissue-specific genes in major depressive disorder and bipolar disorder.	Han et al. (73)	Explore the expression specific characteristics of different areas by systematic analysis of larger samples of brain tissues and determine gene expression patterns and tissue-specific expression profiles between major depressive disorder and bipolar disorder.	Transcriptomic datasets retrieved from the Gene Expression Omnibus (GEO).	Data-guided approach with a weighted gene co-expression network analysis to construct gene co-expression networks for large scale gene expression profiling from various regions of the brain.	Give insights in the tissue-specific functions of various brain regions in the context of psychiatric disorders (MDD and BD). It is a report on functional similarities and specificities between tissues of two psychiatric disorders.
Dissecting psychiatric spectrum disorders by generative embedding.	Brodersen et al. (102)	Examine the feasibility of defining subgroups in psychiatric spectrum disorders by generative embedding.	Functional MRI dataset performing a working memory task.	Theory-driven approach through the use of generative embedding. The researchers used parameter estimates from a dynamic causal model (DCM) of a visual-parietal-prefrontal network to define a model-based feature space for the subsequent application of supervised and unsupervised learning techniques.	This is a proof-of-concept study to examine how model-based clustering could be used to dissect psychiatric spectrum diseases into physiologically defined subgroups, giving foundation to possible implications in the delivery of precision psychiatry. It gives insight into the constraints of a model-guided approach according to its assumptions.
Uncovering social-contextual and individual mental health factors associated with violence via computational inference.	Santamaría-García et al. (103)	Evaluate individual mental health and sociocontextual determinant of violence simultaneously and explore their association to different domains of violence.	Data was taken from a sample of 26,349 ex-members of Colombian illegal armed groups who entered programs of transitional justice for reincorporation into civilian life. They responded to a semi-structured interview designed by the Agency for Reintegration and Normalization.	Combination of theory- and data-driven approaches of examination and analysis of historical records of ex-members of illegal armed groups in Colombia, using deep learning and machine learning methods to identify the most relevant factors associated with domains of violence.	This study investigates the interaction of contextual and individual factors associated with violence in the Colombian context with novel methodologies to take into account historical assessments. Another important aspect of this study is the usage of a combination of theory- and data-driven approaches. This study is not focused in a mental disorder, however it has been weighed the importance of social and individual mental health variables like violence.
Predicting suicide attempts and suicide deaths following outpatient visits using electronic health records.	Simon et al. (157)	Develop and validate models using electronic health records to predict suicide attempt and suicide death following an outpatient visit.	Health care records from seven health systems of 2,960,929 patients.	Data-driven approach to develop prediction models, which were separated between mental health specialty and primary care visits.	This study describes an analysis of a great amount of data across different health care systems. Within the supplementary material, there is a public repository including specifications and code for defining predictor and outcome variables alongside a data dictionary and descriptive statistics for analytic data sets, which impact the reproducibility of the study.
Speech structure links the neural and socio-behavioural correlates of psychotic disorders.	Palaniyappan et al. (53)	Investigate the neural basis and the functional relevance of the structural connectedness of speech samples of subjects with schizophrenia and bipolar disorder.	Clinical assessments of 34 patients with schizophrenia and 22 with bipolar disorder.	Data-driven software to measure graph attributes of connected speech.	This study exemplifies the possibility of establishing a relationship between pathological phenomenology and biological markers. This opens up the possibility of integrating this tool with other computational approaches to achieve a multilevel analysis.

(Continued)

TABLE 1 (Continued)

Title	References	Study aim	Data analyzed	Computational approach	Conclusions/Implications
Pattern of neural responses to verbal fluency shows diagnostic specificity for schizophrenia and bipolar disorder.	Costafreda et al. (109)	Through the usage of the verbal fluency task, the researchers investigated the functional neuroanatomy of executive function in schizophrenia and bipolar disorder. The hypothesis was that the pattern of regional brain responses would correctly identify the diagnosis for each participant at the individual level.	Patients with schizophrenia and bipolar disorder in remission. They were subjected to a clinical assessment and were taken fMRI.	Data-guided approach with the use of machine learning to conduct a pattern classification analysis.	The study highlights the possibility of being able to integrate data from a neurocognitive task and reveal its neurobiological basis to determine precisely diagnostic differences between different clinical entities. It also highlights that the difference between diagnosis comes from degrees of functionality and the limitation of discriminating between them.
Prediction models of functional outcomes for individuals in the clinical high-risk state for psychosis or with recent-onset depression: a multimodal, multisite machine learning analysis.	Koutsouleris et al. (112)	Determine whether predictors associated with social and role functioning can be identified in patients in clinical high-risk states (CHR) for psychosis or with recent-onset depression (ROD) using clinical and imaging-based determinant with machine learning analysis. Assess the geographic, transdiagnostic and prognostic generalizability of machine learning and compare it with human prognostication. Explore sequential prognosis encompassing clinical and combined machine learning.	116 patients in CHR states and 120 patients with ROD.	Data-driven approach using machine learning. Three models of prediction were used (one with clinical variables, one with neuroimaging variables and one integrating the other two).	This study not only explore the predictive model from a data driven approach, but it was also geographically validated. The researchers tested the transferability of the model to other outcomes. It also takes into account the reliability of the inputs which were feeding the model. This study inquires about social factors that drive the personal and socioeconomic burden of psychotic and mood disorders integrating clinical and brain structural data.
Natural language processing of clinical mental health notes may add predictive value to existing suicide risk models.	Levis et al. (113)	Determine if the use of natural language processing (NLP) in psychotherapy note text can provide additional accuracy over currently used suicide prediction models (REACH VET).	Data from the Department of Veterans Affairs (VA) of patients newly diagnosed with PTSD between 2004 and 2013.	Data-driven approach which uses NLP to evaluate unstructured electronic medical records of a sample from de VHA PTSD treatment population.	The method presented in this paper introduces to a dynamic model that helps identify and monitor predictor variables and how they change over time. This gets closer to an ecologically valid tool to asses an individual. This type of approaches on NLP have been used in other pathologies like delirium (158), Alzheimer's disease (159, 160), schizophrenia and others (161, 162).
A machine learning approach using EEG data to predict response to SSRI treatment for major depressive disorder.	Khodayari-Rostamabad et al. (163)	Evaluate the performance of a machine learning methodology based on the pre-treatment electroencephalogram for prediction of response to treatment with SSRI in patients with MDD.	Subjects with MDD derived from a tertiary Mood Disorders Clinic. They were all considered treatment resistant.	Data-driven approach using machine learning to select the most discriminating features from EEG. Then, these features are fed into a classifier based on a mixture factor analysis to give a likelihood value.	This study exemplifies a possible approach to improve treatment in a personalized manner in line with precision psychiatry.
Cross-trial prediction of treatment outcome in depression: a machine learning approach.	Chekroud et al. (119)	Develop an algorithm to assess whether patients will achieve symptomatic remission from a 12-week course of citalopram.	Data was collected from a STAR-D sample.	Data-driven approach using machine learning to identify which variables were most predictive of treatment outcome.	This study determines the possibility of using computational approaches to mine existing clinical trial data to improve on accuracy of risk or treatment response prediction. However, this model only predicts response to specific drugs. There has to be a contextualization of the applicability of the model.

(Continued)

TABLE 1 (Continued)

Title	References	Study aim	Data analyzed	Computational approach	Conclusions/Implications
Gyri-precise head model of transcranial direct current stimulation: improved spatial focality using a ring electrode versus conventional rectangular pad.	Datta et al. (123)	Compare the focality of conventional rectangular-pad stimulation with ring electrode configuration using a MRI-derived head model.	Models of two electrode configurations.	Use of a head model to predict relative spatial focality and the influence of tissue geometry/conductivity.	This study demonstrates a way of translate computational models of variables associated with treatments such as direct current stimulation to clinical applications through the design and optimization of treatment variables.
Effects of modifying the electrode placement and pulse width on cognitive side effects with unilateral ECT: a pilot randomized controlled study with computational modelling.	Martin et al. (124)	Determine if the frontoparietal placement of electrodes improves retrograde memory outcomes compared to temporoparietal placement.	Patients recruited from a single hospital in Sydney.	Computational model (164) was used in a subset of participants to determine if higher levels of stimulation in regions of interest would be related to worse or better cognitive outcomes.	This study gives an example of how data from computational models could be integrated to results from clinical investigations to individualize treatment options such as DCS.
Patient-specific analysis of the volume of tissue activated during deep brain stimulation.	Butson et al. (125)	Develop and test a methodology that would enable prediction and visualization of the volume of axonal tissue activated during DBS.	One patient with Parkinson's disease.	Patient-specific model of STN DBS for PD and the VTAs. This model was constructed from 3D brain atlas that was warped to the patient MRI using a non-linear warping algorithm. The electrical and biophysical models rely on finite element models.	This model integrates anatomical, electrical, and biophysical representation of DBS. It also integrates simulation data with clinical data from subject. The limitation of this model is the evaluation of only one patient.
Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity.	Finn et al. (126)	Determine if functional connectivity profiles can act as an identifying fingerprint capable of identifying an individual from a set of connectivity profiles.	Data collected from the Human Connectome Project.	Data-driven approach using a group-wise spectral clustering algorithm for the definition of networks capable of being compared to each other. This correlation was made through the use of whole-brain connectivity matrix.	This study gives the foundation for novel test inferences about functional brain organization can relate to distinct behavioral phenotypes. The discriminating power evidenced in this study is partly the result of the relatively long period of time of follow-up. This can be integrated in frameworks like RDoC. It also gives the base for neuroimaging studies which rely on single subjects, beyond population-level studies.
Linked dimensions of psychopathology and connectivity in functional brain networks.	Xia et al. (127)	Identify brain-based dimensions of psychopathology.	Datasets taken from the Philadelphia Neurodevelopmental Cohort (PNC).	Data-driven approach based on sparse canonical correlation analysis.	This study uses network theory to construct patterns of functional connectivity, which could be linked to transdiagnostic dimensions of psychopathology. In this study, these patterns displayed developmental and sex differences. This in turn tackles the problems of comorbidity and heterogeneity previously discussed in this article.

(Continued)

TABLE 1 (Continued)

Title	References	Study aim	Data analyzed	Computational approach	Conclusions/Implications
Origin of cyclicity in bipolar disorders: a computational approach.	Goldbeter (131)	Evaluate a model for bipolar disorders based on mutual inhibition of two putative neural circuits governing the affective syndromes.	Mathematical model based on reciprocal inhibition.	Theory-driven approach of a mathematical model to predict the cyclicity of bipolar disorders. This model is based on a phenomenological model.	This article gives an example of translating a phenomenological level to mathematical terms in order to explain and predict a characteristic of a phenomenon (cyclicity of bipolar disorders).
Computational mechanism of effort and reward decisions in patients with depression and their association with relapse after antidepressant discontinuation.	Berwian et al. (132)	Establish whether the decision to invest effort for rewards represents a persistent depression process after remission.	Sample of patients in a Swiss and German university setting.	Theory-driven approach where a generative computational model was used to represent the putative computations of the behavioral pattern.	This study explores a computational model for effortful behavior applied in a sample of patients with depression. This gives a straightforward manner to assess this behavioral feature and find associations that are important form a prognosis and treatment perspective. Nonetheless, this study has limitations from a replicability perspective.
Making working memory work: a computational model of learning in the prefrontal cortex and basal ganglia.	O'reilly and Frank (133)	Presentation of a computational model of working memory based on the prefrontal cortex and basal ganglia.	The 1-2-AX task.	Theory-driven approach which uses a reinforcement learning mechanism.	This paper describes how a theory-driven model is constructed from data previously acquired which is integrated to elucidate a specific process.
Towards a computational psychiatry of juvenile obsessive-compulsive disorder.	Loosen and Hauser (134)	Review computational, neuropsychological and neural alterations in juvenile OCD. Link these findings to adult OCD. Establish a neurocomputational framework that illustrates the development of symptoms in the context of juvenile OCD.	Narrative review.	Theory-driven approach based on a narrative review of computational, neuropsychological and neural alterations in juvenile OCD. The framework proposed is based on a meta-controller with different rates of maturation of complex systems.	This study describes a proposition of a theory-driven model for the development of obsessive symptoms. However, this model is only speculative and requires further investigation to be validated. It highlights the importance of the *a priori* knowledge to construct the model and the dependance on inputs to determine the strength of the model.
Adaptive current-flow models of ECT: explaining individual static impedance, dynamic impedance, and brain current density.	Unal et al. (138)	Examine the relations between the physical properties of the ECT stimulus, patient head anatomy, and patient-specific impedance to the passage of current.	Clinical data from a trial series.	Theory-driven approach. The researchers develop an individualize (MRI-derived) finite element method (FEM) to model transcranial electrical stimulation with dynamical changes in tissue conductivity.	This model gives the opportunity of studying parameters that have been proposed as important factors in the therapeutic response (165), but they are difficult to study under a "normal" clinical study.

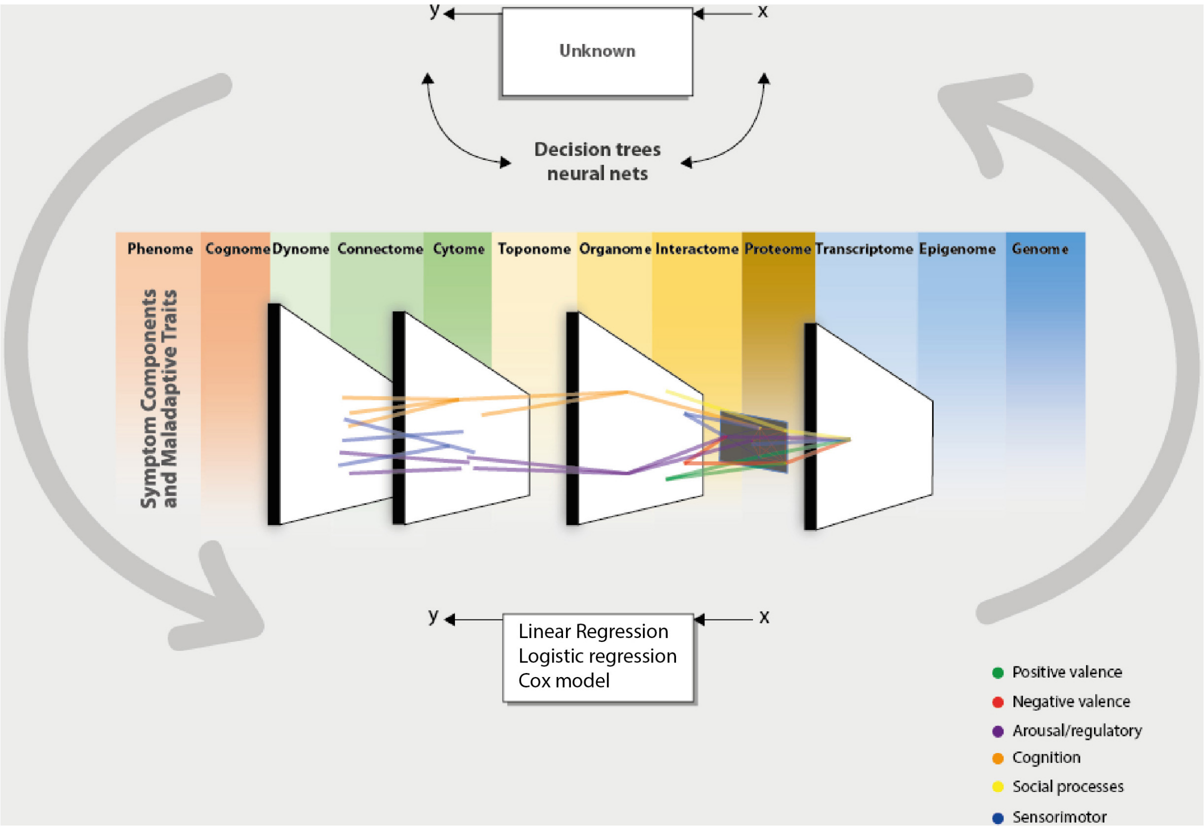


FIGURE 1
Computational psychiatry aids the clinician and the researcher in integrating data from different sources of information, which could be taken from the omics perspective. This integration is made possible by complementing the data modeling culture using the algorithmic modeling culture proposed by Breiman (43). This permits the validation of models or data that can be measured by predictive accuracy. By taking these inputs and processing them through a computational system (algorithm), one could present data-driven or theory-driven responses to clinical and research questions. This enables us to bring forward an integrative and cohesive framework associated with others. The network theory can integrate the different units of analysis (scale level) of a phenomenon or give a cohesive picture of the interaction between different domains. And in turn, this could give us a more precise phenotype to arrive at a dimensional conception (HiTOP). These computational approaches to understanding psychiatry represent the brain's functioning [phantastic organ (20)]. In other words, using computational approaches to comprehend psychiatry mimics the normal functioning of the statistical machine we call the brain.

disorders. It recognizes that mental disorders are multicausal, mediated by biology (brain). In addition, the RDoC is structured as a matrix with different units of analysis, which are grouped into research domains. These domains are viewed longitudinally, influenced by neurodevelopment and the context in which they are imbued. Computational psychiatry, then, introduces itself as a great tool in this type of initiative, aligning with its principles, since it allows to appreciate of shared mechanisms between cognitive alterations, psychopathological domains, and disorders (59), achieving integration between the different levels of analysis (units of analysis and domains). It does this by finding objective, observable, and measurable characteristics organized into taxonomies outside current nosology (25), achieving a more solid basis for neurobiological research.

With this initiative, it has been possible to see that in most mental disorders, there is an overlap between neural circuits in which the processing of threats (amygdala, hippocampus, orbitofrontal cortex, and ventromedial prefrontal cortex), rewards (amygdala, ventral tegmental area, locus coeruleus, and nucleus accumbens) and perception of stimuli (thalamus, sensory cortex, and inferior frontal gyrus) are counted (25). This suggests that mental disorders may be due to different modes of dysregulation of control processes. That is a different dynamic system. These altered processes can occur from the cellular and molecular level to the level of circuits. And this generates a greater difficulty since the alterations will only vary qualitatively but quantitatively. This, at a practical level, limits the possibility of using only clinical judgment to determine these nuances. Again, the problem with these ambiguous cases, which are the rule and not the exception in psychiatry, is highlighted by the lack of persistence in diagnosis given to a person over time and the problem of comorbidity and heterogeneity (147).

However, it does present guidelines that can be a response to the criticism previously mentioned of nosology and psychiatric research based on it, as well as a bridge for using computational models to the approach of multidimensional and hierarchical organization of mental functions, the non-linear dynamic interaction between the components of the system and its heterogeneity. Thus, computational psychiatry aligns with one of the objectives of the RDoC, which is to improve the accuracy of the phenotypes and their alignment with highly plausible biological and cognitive models based on experimental settings used in neuroscience research applied to psychiatry (87).

Nonetheless, this is one of many models which have risen to deal with the limitations and constraints previously described.

The HiTOP is a data-driven, hierarchically based organization of psychopathology (86). It conceptualizes psychopathology as a set of dimensions organized into increasingly broad, transdiagnostic spectra. This is made by using factor analysis between different symptoms to generate a taxonomy of mental disorders. In these scenarios, the computational models would aid in determining these psychopathologic patterns using path analyses in clinical datasets. Lastly, they would also be helpful in establishing psychopathologic patterns taking into account their context (social influences and contextual factors).

Another alternative is the network theory based on pattern analysis, similar to computational psychiatry. To construct these networks, one has to analyze a significant amount of data that can capture cohesion, coherence, and patterns of synchrony (148). In this sense, computational psychiatry dialogues with the network approach both require massive data processing to formulate theoric models.

As previously discussed, HiTOP is another proposed model for this endeavor. It was constructed through factor analysis and latent class analysis to organize psychopathology according to the natural covariance structure between symptoms, maladaptive behaviors, and traits (61, 86). This model focuses on the psychopathological level remaining agnostic to the underlying phenotypes encountered. Moreover, it can be a tool to aid RDoC-informed research by providing psychometrically valid data to reach more robust psychiatric phenotypes (149), and in doing so, it can ameliorate the computational models used in clinical and research fields.

General limitations

One of the limitations that must be considered in the explanatory models is that the data previously collected empirically may contain significant biases that prevent distinguishing between different hypotheses of the mechanisms that generate psychiatric dysfunctions. For this reason, it is of the utmost importance to recognize parameters that allow discriminating between models (59). On the other hand, for data-driven approaches, the clinical datasets from which one can take the information are limited in data quality, organization and accessibility, making it difficult to get the data for the machine learning algorithms (100, 150).

Another limitation is inherent to mental disorders since they usually present dysfunctions or deficits that are generalized, shared by many disorders, and only differentiated at a quantitative level (59), so a large enough sample must identify these differences. This limitation can be overcome by the formation of consortiums like the ones developed for genomics studies and others (151).

Still, another limitation is that, in most cases of mental disorders, the brain regions or alterations underlying a particular dysfunction have not been accurately determined. However, reverse-engineering strategies can overcome this automatically, seeking to identify physical and biological laws through data (25). However, this raises another problem because these structures can be purely mathematical entities that do not have a basis in biological structures.

In addition to this, neurobiological models describe data as unreliable, meaning that the probability of error in the model must be quantified (25). These errors are critical in these models, where many interrelated variables spread that error to different parts of the system. Typically, this type of error is controlled by increasing the sample size; however, in this case, it would worsen the problem because it could result in inaccurate models with statistical significance (25). Moreover, brain processing is non-linear, having complex interrelationships, amplifying, or decreasing the noise of the inputs. This causes linear regressions to lose their significance. In neurobiological responses, various processes like serial signaling processes, thresholds, filters, saturation, feedback, etc. All of these are non-linear and, therefore, more difficult to describe.

Also, using Bayes' theorem to choose the best model will often lead to models that do not describe the best generative model. Therefore, one should always validate the model (35) and always keep in mind the possibility of finding better models.

Conclusion

Computational psychiatry can be a tool for understanding mental health. This involves a great effort, which requires the articulation of multiple disciplines and different levels of analysis. Therefore computational psychiatry could become a high point and central to attempts such as RDoC or ROAMER (Roadmap for Mental Health Research in Europe) (152) to achieve a better conception of both mental disorders and mental health, with the articulation with other models like HiTOP.

To achieve this, it is necessary to overcome previously evidenced obstacles such as heterogeneity and comorbidity, together with the acceptance and use of the complexity of these systems with non-linear dynamics, making use of tools that allow us to understand it in a way in both biological and psychological reductionist perspectives are not given. In addition, the opportunity opens up to begin the study, articulation, and integration with factors that modulate the presentation and prognosis of mental disorders but that are left to the context and have been covered only tangentially, as are social processes such as violence, abuse or forced displacement. The development of research capacity achieves a better assessment of the needs for the care of the population, increasing knowledge about the effectiveness of different interventions and creating a critical mass that is essential for the development of the scientific debate on various topics in mental health (153). In addition, this theoretical framework model how the subject acquires and transforms their internal cognitive processes to give rise to their behavioral responses, which are observed in clinical practice (15).

However, to accomplish all these promises, several limitations must be considered. The necessity for not only replicating the results of different investigations arises with the need for reproducible investigations to tackle the falsifiability problem. In this same direction, with the growth of analytic power, the possibility of finding associations that are not significant or valid also increases. So, the validation of these models is yet another fundamental aspect that must be tackled by researchers.

Finally, computational psychiatry would allow us to provide better care for mental health problems in primary care. This considers that the burden of patients increases with poor support for the number of professionals in mental health. Then, the models given by computational psychiatry would allow the specialist to have better visualization and contextualization of each patient's specific case considering multiple factors that often cannot be given enough weight due to restrictions. Also, data-based computational models allow predictions or diagnostics, and these responses can be better adjusted to the context of each country and can be free or require low investments.

The promises are manifold, but their success depends on their applicability and the possibility of generating translational knowledge (154), **Figure 1** proposes a framework to arrive at this result.

Author contributions

JC conceived the general idea of the review and summarized the data found through the search, which was edited by HS-G. Both authors developed the search criteria to complete the review, analyzed the data found, and revised it critically to arrive at the approved version to be published.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. McCrone J. Friston's theory of everything. *Lancet Neurol.* (2022) 21:494.

2. Friston K. The free-energy principle: a unified brain theory? *Nat Rev Neurosci.* (2010) 11:127–38.

3. Douven I, Schupbach J. Probabilistic alternatives to bayesianism: the case of explanationism. *Front Psychol.* (2015) 6:459. doi: 10.3389/fpsyg.2015.00459

4. Phillips J, Frances A, Cerullo M, Chardavoyne J, Decker H, First M, et al. The six most essential questions in psychiatric diagnosis: a pluralogue. Part 4: general conclusion. *Philos Ethics Humanit Med.* (2012) 7:14. doi: 10.1186/1747-5341-7-14

5. Phillips J, Frances A, Cerullo M, Chardavoyne J, Decker H, First M, et al. The six most essential questions in psychiatric diagnosis: a pluralogue part 1: conceptual and definitional issues in psychiatric diagnosis. *Philos Ethics Humanit Med.* (2012) 7:3. doi: 10.1186/1747-5341-7-3

6. Phillips J, Frances A, Cerullo M, Chardavoyne J, Decker H, First M, et al. The six most essential questions in psychiatric diagnosis: a pluralogue part 3: issues of utility and alternative approaches in psychiatric diagnosis. *Philos Ethics Humanit Med.* (2012) 7:9. doi: 10.1186/1747-5341-7-9

7. Salessi S. Aporia of power: on the crises, science, and internal dynamics of the mental health field. *Eur J Philos Sci.* (2017) 7:175–200.

8. Aboraya A, France C, Young J, Curci K, Lepage J. The validity of psychiatric diagnosis revisited: the clinician's guide to improve the validity of psychiatric diagnosis. *Psychiatry.* (2005) 2:48–55.

9. Venkatasubramanian G, Keshavan M. Biomarkers in psychiatry-a critique. *Ann Neurosci.* (2016) 23:3–5.

10. Mcgorry P, Keshavan M, Goldstone S, Amminger P, Allott K, Berk M, et al. Biomarkers and clinical staging in psychiatry. *World Psychiatry.* (2014) 13:211–23.

11. Kirmayer L, Crafa D. What kind of science for psychiatry? *Front Hum Neurosci.* (2014) 8:435. doi: 10.3389/fnhum.2014.00435

12. Sadock B, Kaplan H, Ruiz P. *Kaplan & sadock's comprehensive textbook of psychiatry*. 10th ed. Philadelphia: Lippincott Williams & Wilkins (2017).

13. Peled A. Neuroscientific psychiatric diagnosis. *Med Hypotheses.* (2009) 73:220–9.

14. Mäki-Marttunen T, Kaufmann T, Elvsåshagen T, Devor A, Djurovic S, Westlye L, et al. Biophysical psychiatry—how computational neuroscience can help to understand the complex mechanisms of mental disorders. *Front Psychiatry.* (2019) 10:534. doi: 10.3389/fpsyt.2019.00534

15. Stephan K, Mathys C. Computational approaches to psychiatry. *Curr Opin Neurobiol.* (2014) 25:85–92.

16. Ito T, Hearne L, Mill R, Cocuzza C, Cole M. Discovering the computational relevance of brain network organization. *Trends Cogn Sci.* (2020) 24:25–38.

17. Trappenberg T. *Fundamentals of computational neuroscience*. Oxford: Oxford University Press (2010).

18. Shagrir O. Marr on computational-level theories. *Philos Sci.* (2010) 77:477–500.

19. van den Bos W, Bruckner R, Nassar M, Mata R, Eppinger B. Computational neuroscience across the lifespan: promises and pitfalls. *Dev Cogn Neurosci.* (2018) 33:42–53. doi: 10.1016/j.dcn.2017.09.008

20. Friston K, Enno Stephan K, Montague R, Dolan R. Computational psychiatry: the brain as a phantastic organ. *Lancet Psychiatry.* (2014) 1:148–58. doi: 10.1016/S2215-0366(14)70275-5

21. Rutledge R, Chekroud A, Jm Huys Q. Machine learning and big data in psychiatry: toward clinical applications. *Curr Opin Neurobiol.* (2019) 55:152–9. doi: 10.1016/j.conb.2019.02.006

22. Tai A, Albuquerque A, Carmona N, Subramanieapillai M, Cha D, Sheko M, et al. Machine learning and big data: implications for disease modeling and therapeutic discovery in psychiatry. *Artif Intell Med.* (2019) 99:101704. doi: 10.1016/j.artmed.2019.101704

23. Bzdok D, Meyer-Lindenberg A. Machine learning for precision psychiatry: opportunities and challenges. *Biol Psychiatry Cogn Neurosci Neuroimaging.* (2018) 3:223–30.

24. Gillan C, Whelan R. What big data can do for treatment in psychiatry. *Curr Opin Behav Sci.* (2017) 18:34–42.

25. Mujica-Parodi L, Strey H. Making sense of computational psychiatry. *Int J Neuropsychopharmacol.* (2020) 23:339–47.

26. Weston A. Décima. In: Malem J editor. *Las claves de la argumentación*. Barcelona: Ariel (2006).

27. Machado-Vieira R. Tracking the impact of translational research in psychiatry: state of the art and perspectives. *J Transl Med.* (2012) 10:175. doi: 10.1186/1479-5876-10-175

28. Allsopp K, Read J, Corcoran R, Kinderman P. Heterogeneity in psychiatric diagnostic classification. *Psychiatry Res.* (2019) 279:15–22.

29. Newman L, Polk T. The computational cognitive neuroscience of learning and memory: principles and models. *Adv Psychol.* (2008) 139:77–99.

30. Feczko E, Miranda-Dominguez O, Marr M, Graham A, Nigg J, Fair D. The heterogeneity problem: approaches to identify psychiatric subtypes. *Trends Cogn Sci.* (2019) 23:584–601.

31. Nomi J. Regression models for characterizing categorical-dimensional brain-behavior relationships in clinical populations. *Biol Psychiatry Cogn Neurosci Neuroimaging.* (2019) 4:419–20. doi: 10.1016/j.bpsc.2019.03.008

32. Santamaría-García H, Reyes P, García A, Baéz S, Martinez A, Santacruz J, et al. First symptoms and neurocognitive correlates of behavioral variant frontotemporal dementia. *J Alzheimers Dis.* (2016) 54:957–70.

33. Donnelly-Kehoe P, Pascariello G, García A, Hodges J, Miller B, Rosen H, et al. Robust automated computational approach for classifying frontotemporal neurodegeneration: multimodal/multicenter neuroimaging. *Alzheimers Dement.* (2019) 11:588–98. doi: 10.1016/j.dadm.2019.06.002

34. Abrevaya S, Fittipaldi S, Garcia A, Dottori M, Santamaria-Garcia H, Birba A, et al. At the heart of neurological dimensionality: cross-nosological and multimodal cardiac interoceptive deficits. *Psychosom Med.* (2020) 82:850–61. doi: 10.1097/PSY.0000000000000868

35. Adams R, Huys Q, Roiser J. Computational psychiatry: towards a mathematically informed understanding of mental illness. *J Neurol Neurosurg Psychiatry.* (2016) 87:53–63.

36. Maito M, Santamaría-García H, Moguilner S, Possin K, Godoy M, Avila-Funes J, et al. Classification of Alzheimer's disease and frontotemporal dementia using routine clinical and cognitive measures across multicentric underrepresented samples: a cross sectional observational study. *Lancet Reg Health Am.* (2023) 17:100387. doi: 10.1016/j.lana.2022.100387

37. Parmigiani G, Barchielli B, Casale S, Mancini T, Ferracuti S. The impact of machine learning in predicting risk of violence: a systematic review. *Front Psychiatry.* (2022) 13:1015914. doi: 10.3389/fpsyt.2022.1015914

38. Clark A. Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behav Brain Sci.* (2013) 36:181–204.

39. Friston K. Computational psychiatry: from synapses to sentience. *Mol Psychiatry* (2022) 28:1–13. doi: 10.1038/s41380-022-01743-z

40. Knill D, Pouget A. The bayesian brain: the role of uncertainty in neural coding and computation. *Trends Neurosci.* (2004) 27:712–9.

41. Thá F, da Silveira E, da Silveira T. The hysterical symptom: a proposal of articulation of the freudian theory and the bayesian account. *Neuropsychoanalysis.* (2021) 23:83–95. doi: 10.1080/15294145.2021.1999845

42. Hardcastle V, Hardcastle K. Marr's levels revisited: understanding how brains break. *Top Cogn Sci.* (2015) 7:259–73. doi: 10.1111/tops.12130

43. Breiman L. Statistical modeling: the two cultures. *Stat Sci.* (2001) 16:199–231.

44. Huys Q, Maia T, Frank M. Computational psychiatry as a bridge from neuroscience to clinical applications. *Nat Neurosci.* (2016) 19:404–13.

45. Mathys C, Daunizeau J, Friston K, Stephan K. A bayesian foundation for individual learning under uncertainty. *Front Hum Neurosci.* (2011) 5:39. doi: 10.3389/fnhum.2011.00039

46. Huys Q. Advancing clinical improvements for patients using the theory-driven and data-driven branches of computational psychiatry. *JAMA Psychiatry.* (2018) 75:225–6. doi: 10.1001/jamapsychiatry.2017.4246

47. Bastos A, Usrey W, Adams R, Mangun G, Fries P, Friston K. Canonical microcircuits for predictive coding. *Neuron.* (2012) 76:695–711.

48. Mikulasch F, Rudelt L, Wibral M, Priesemann V. Where is the error? Hierarchical predictive coding through dendritic error computation. *Trends Neurosci.* (2023) 46:45–59. doi: 10.1016/j.tins.2022.09.007

49. Cumming G. Heterarchies: reconciling networks and hierarchies. *Trends Ecol Evol.* (2016) 31:622–32. doi: 10.1016/j.tree.2016.04.009

50. Bechtel W. Resituating cognitive mechanisms within heterarchical networks controlling physiology and behavior. *Theory Psychol.* (2019) 29:620–39.

51. Cohen A, Elvevåg B. Automated computerized analysis of speech in psychiatric disorders. *Curr Opin Psychiatry.* (2014) 27:203–9.

52. Mota N, Vasconcelos N, Lemos N, Pieretti A, Kinouchi O, Cecchi G, et al. Speech graphs provide a quantitative measure of thought disorder in psychosis. *PLoS One.* (2012) 7:e34928. doi: 10.1371/journal.pone.0034928

53. Palaniyappan L, Mota N, Oowise S, Balain V, Copelli M, Ribeiro S, et al. Speech structure links the neural and socio-behavioural correlates of psychotic disorders. *Prog Neuropsychopharmacol Biol Psychiatry.* (2019) 88:112–20. doi: 10.1016/j.pnpbp.2018.07.007

54. Mota N, Copelli M, Ribeiro S. Thought disorder measured as random speech structure classifies negative symptoms and schizophrenia diagnosis 6 months in advance. *NPJ Schizophr.* (2017) 3:18. doi: 10.1038/s41537-017-0019-3

55. Mota N, Sigman M, Cecchi G, Copelli M, Ribeiro S. The maturation of speech structure in psychosis is resistant to formal education. *NPJ Schizophr.* (2018) 4:25. doi: 10.1038/s41537-018-0067-3

56. Malcorra B, Mota N, Weissheimer J, Schilling L, Wilson M, Hübner L. Low speech connectedness in alzheimer's disease is associated with poorer semantic memory performance. *J Alzheimers Dis.* (2021) 82:905–12. doi: 10.3233/JAD-210134

57. Parola A, Lin J, Simonsen A, Bliksted V, Zhou Y, Wang H, et al. Speech disturbances in schizophrenia: assessing cross-linguistic generalizability of NLP automated measures of coherence. *Schizophr Res.* (2022). doi: 10.1016/j.schres.2022.07.002 [Epub ahead of print].

58. Clarke N, Foltz P, Garrard P. How to do things with (thousands of) words: computational approaches to discourse analysis in Alzheimer's disease. *Cortex.* (2020) 129:446–63. doi: 10.1016/j.cortex.2020.05.001

59. Bennett D, Silverstein S, Niv Y. The two cultures of computational psychiatry. *JAMA Psychiatry.* (2019) 76:563–4.

60. Vasil J, Badcock P, Constant A, Friston K, Ramstead M. A world unto itself: human communication as active inference. *Front Psychol.* (2020) 11:417. doi: 10.3389/fpsyg.2020.00417

61. Kotov R, Waszczuk M, Krueger R, Forbes M, Watson D, Clark L, et al. The hierarchical taxonomy of psychopathology (HiTOP): a dimensional alternative to traditional nosologies. *J Abnorm Psychol.* (2017) 126:454–77.

62. Borsboom D. A network theory of mental disorders. *World Psychiatry.* (2017) 16:5–13.

63. Rall W. Branching dendritic trees and motoneuron membrane resistivity. *Exp Neurol.* (1959) 1:491–527.

64. Hodgkin A, Huxley AF. A quantitative description of membrane current and its application to conduction and excitation in nerve. *J Physiol.* (1952) 117:500–44.

65. Townsley K, Brennand K, Huckins L. Massively parallel techniques for cataloguing the regulome of the human brain. *Nat Neurosci.* (2020) 23:1509–21. doi: 10.1038/s41593-020-00740-1

66. Grove J, Ripke S, Als T, Mattheisen M, Walters R, Won H, et al. Identification of common genetic risk variants for autism spectrum disorder. *Nat Genet.* (2019) 51:431–44.

67. Guo W, Yu D, Davis L, Ripke S, Shugart Y, Arnold P, et al. Revealing the complex genetic architecture of obsessive- compulsive disorder using meta-analysis: international obsessive compulsive disorder foundation genetics collaborative (IOCDF- GC) and OCD collaborative genetics association studies (OCGAS). *Mol Psychiatry.* (2018) 23:1181–8. doi: 10.1038/mp.2017.154

68. Demontis D, Walters R, Martin J, Mattheisen M, Als T, Agerbo E, et al. Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder. *Nat Genet.* (2019) 51:63–75.

69. The Schizophrenia Working Group of the Psychiatric Genomics Consortium, Ripke S, Walters J, O'Donovan M. Mapping genomic loci prioritises genes and implicates synaptic biology in schizophrenia. *medRxiv.* [Preprint]. (2020). doi: 10.1101/2020.09.12.20192922

70. Jansen I, Savage J, Watanabe K, Bryois J, Williams D, Steinberg S, et al. Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nat Genet.* (2019) 51:404–13.

71. Watson H, Yilmaz Z, Thornton L, Hübel C, Coleman J, Gaspar H, et al. Genome-wide association study identifies eight risk loci and implicates metabo-psychiatric origins for anorexia nervosa. *Nat Genet.* (2019) 51:1207–14. doi: 10.1038/s41588-019-0439-2

72. Browne F, Wang H, Zheng H. A computational framework for the prioritization of disease-gene candidates. *BMC Genomics.* (2015) 16:S2. doi: 10.1186/1471-2164-16-S9-S2

73. Han M, Yuan L, Huang Y, Wang G, Du C, Wang Q, et al. Integrated co-expression network analysis uncovers novel tissue-specific genes in major depressive disorder and bipolar disorder. *Front Psychiatry.* (2022) 13:980315. doi: 10.3389/fpsyt.2022.980315

74. Kesner R, Rolls E. A computational theory of hippocampal function, and tests of the theory: new developments. *Neurosci Biobehav Rev.* (2015) 48:92–147.

75. Larner A. Transient global amnesia: model, mechanism, hypothesis. *Cortex.* (2022) 149:137–47.

76. Murray J, Demirtaş M, Anticevic A. Biophysical modeling of large-scale brain dynamics and applications for computational psychiatry. *Biol Psychiatry Cogn Neurosci Neuroimaging.* (2018) 3:777–87. doi: 10.1016/j.bpsc.2018.07.004

77. Wang B, Guo M, Pan T, Li Z, Li Y, Xiang J, et al. Altered higher-order coupling between brain structure and function with embedded vector representations of connectomes in schizophrenia. *Cereb Cortex.* (2022). doi: 10.1093/cercor/bhac432 [Epub ahead of print].

78. Thomas P, Leow A, Klumpp H, Phan K, Ajilore O. Network diffusion embedding reveals transdiagnostic subnetwork disruption and potential treatment targets in internalizing psychopathologies. *Cereb Cortex.* (2022) 32:1823–39. doi: 10.1093/cercor/bhab314

79. Myles L. The emerging role of computational psychopathology in clinical psychology. *Mediterr J Clin Psychol.* (2021) 9:1–7.

80. Stein D, Phillips K, Bolton D, Fulford K, Sadler J, Kendler K. What is a mental/psychiatric disorder? from DSM-IV to DSM-V. *Psychol Med.* (2010) 40:1759–65.

81. Nemeroff C, Weinberger D, Rutter M, MacMillan H, Bryant R, Wessely S, et al. DSM-5: a collection of psychiatrist views on the changes, controversies, and future directions. *BMC Med.* (2013) 11:202. doi: 10.1186/1741-7015-11-202

82. Clark L, Cuthbert B, Lewis-Fernández R, Narrow W, Reed G. Three approaches to understanding and classifying mental disorder: ICD-11, DSM-5, and the national institute of mental health's research domain criteria (RDoC). *Psychol Sci Public Interest.* (2017) 18:72–145.

83. Andreasen N. DSM and the death of phenomenology in America: an example of unintended consequences. *Schizophr Bull.* (2007) 33:108–12. doi: 10.1093/schbul/sbl054

84. Phillips J, Frances A, Cerullo M, Chardavoyne J, Decker H, First M, et al. The six most essential questions in psychiatric diagnosis: a pluralogue part 2: issues of conservatism and pragmatism in psychiatric diagnosis. *Philos Ethics Humanit Med.* (2012) 7:8. doi: 10.1186/1747-5341-7-8

85. Sperandeo R, Mosca LL, Cioffi V, Davide A, Sarno D, Iennaco D, et al. Complexity in the narration of the self. In: *Proceedings of the 10th IEEE International Conference on Cognitive Infocommunications*. (2019) 445–50.

86. Ruggero C, Kotov R, Hopwood C, First M, Clark L, Skodol A, et al. Integrating the hierarchical taxonomy of psychopathology (HiTOP) into clinical practice. *J Consult Clin Psychol.* (2019) 87:1069–84.

87. Hitchcock P, Fried E, Frank M. Computational psychiatry needs time and context. *Annu Rev Psychol.* (2022) 73:243–70.

88. Eyre M, Fitzgibbon S, Ciarrusta J, Cordero-Grande L, Price A, Poppe T, et al. The developing human connectome project: typical and disrupted perinatal functional connectivity. *Brain.* (2021) 144:2199–213.

89. Calkins M, Merikangas K, Moore T, Burstein M, Behr A, Satterthwaite T, et al. Deep phenotyping collaborative. *J Child Psychiatry.* (2016) 56:1356–69.

90. Kiddle B, Inkster B, Prabhu G, Moutoussis M, Whitaker K, Bullmore E, et al. Cohort profile: the NSPN 2400 cohort: a developmental sample supporting the wellcome trust neuro science in psychiatry network. *Int J Epidemiol.* (2018) 47:18–9g. doi: 10.1093/ije/dyx117

91. Somerville L, Bookheimer S, Buckner R, Burgess G, Curtiss S, Dapretto M. The lifespan human connectome project in development: a large-scale study of brain connectivity development in 5–21 year olds. *Neuroimage.* (2018) 183:456–68. doi: 10.1016/j.neuroimage.2018.08.050

92. Morgan S, White S, Bullmore E, Vértes P. A network neuroscience approach to typical and atypical brain development. *Biol Psychiatry Cogn Neurosci Neuroimaging.* (2018) 3:754–66. doi: 10.1016/j.bpsc.2018.03.003

93. Hyde L. Developmental psychopathology in an era of molecular genetics and neuroimaging: a developmental neurogenetics approach. *Dev Psychopathol.* (2015) 27:587–613. doi: 10.1017/S0954579415000188

94. Palminteri S, Wyart V, Koechlin E. The importance of falsification in computational cognitive modeling. *Trends Cogn Sci.* (2017) 21:425–33. doi: 10.1016/j.tics.2017.03.011

95. Chandler C, Foltz P, Elvevåg B. Using machine learning in psychiatry: the need to establish a framework that nurtures trustworthiness. *Schizophr Bull.* (2020) 46:11–4. doi: 10.1093/schbul/sbz105

96. Peng R. Reproducible research and biostatistics. *Biostatistics.* (2009) 10:405.

97. Munafò M, Nosek B, Bishop D, Button K, Chambers C, du Sert N, et al. A manifesto for reproducible science. *Nat Hum Behav.* (2017) 1:0021.

98. Wagenmakers E, Borsboom D, Verhagen J, Kievit R, Bakker M, Cramer A, et al. The meaning of "significance" for different types of research [Translated and Annotated. *Acta Psychol.* (2014) 148:188–94.

99. American Statistical Association,. *Ethical guidelines for statistical practice.* Boston: American Statistical Association (2018).

100. Hauser T, Skvortsova V, de Choudhury M, Koutsouleris N. The promise of a model-based psychiatry: building computational models of mental ill health. *Lancet Digit Health.* (2022) 4:e816–28. doi: 10.1016/S2589-7500(22)00152-2

101. Friston K, Preller K, Mathys C, Cagnan H, Heinzle J, Razi A, et al. Dynamic causal modelling revisited. *Neuroimage.* (2019) 199:730–44.

102. Brodersen K, Deserno L, Schlagenhauf F, Lin Z, Penny W, Buhmann J, et al. Dissecting psychiatric spectrum disorders by generative embedding. *Neuroimage Clin.* (2014) 4:98–111. doi: 10.1016/j.nicl.2013.11.002

103. Santamaría-García H, Baez S, Aponte-Canencio D, Pasciarello G, Donnelly-Kehoe P, Maggiotti G, et al. Uncovering social-contextual and individual mental health factors associated with violence via computational inference. *Patterns.* (2021) 2:100176. doi: 10.1016/j.patter.2020.100176

104. Campo-Arias A, Herazo E, Reyes-Rojas M. Cultural psychiatry: beyond DSM-5. *Rev Colomb Psiquiatr.* (2021) 50:138–45.

105. Kapur S, Phillips A, Insel T. Why has it taken so long for biological psychiatry to develop clinical tests and what to do about it? *Mol Psychiatry.* (2012) 17:1174–9. doi: 10.1038/mp.2012.105

106. Turecki G, Brent D, Gunnell D, O'Connor R, Oquendo M, Pirkis J, et al. Suicide and suicide risk. *Nat Rev Dis Primers.* (2019) 5:74.

107. Fazel S, Runeson B. Suicide. *N Engl J Med.* (2020) 382:266–74.

108. Neuner T, Schmid R, Wolfersdorf M, Spießl H. Predicting inpatient suicides and suicide attempts by using clinical routine data? *Gen Hosp Psychiatry.* (2008) 30:324–30.

109. Costafreda S, Fu C, Picchioni M, Toulopoulou T, McDonald C, Kravariti E, et al. Pattern of neural responses to verbal fluency shows diagnostic specificity for schizophrenia and bipolar disorder. *BMC Psychiatry.* (2011) 11:18. doi: 10.1186/1471-244X-11-18

110. Mota N, Furtado R, Maia P, Copelli M, Ribeiro S. Graph analysis of dream reports is especially informative about psychosis. *Sci Rep.* (2014) 4:3691. doi: 10.1038/srep03691

111. Hyman S. The diagnosis of mental disorders: the problem of reification. *Annu Rev Clin Psychol.* (2010) 6:155–79.

112. Koutsouleris N, Kambeitz-Ilankovic L, Ruhrmann S, Rosen M, Ruef A, Dwyer D, et al. Prediction models of functional outcomes for individuals in the clinical high-risk state for psychosis or with recent-onset depression: a multimodal, multisite machine learning analysis. *JAMA Psychiatry.* (2018) 75:1156–72.

113. Levis M, Leonard Westgate C, Gui J, Watts B, Shiner B. Natural language processing of clinical mental health notes may add predictive value to existing suicide risk models. *Psychol Med.* (2021) 51:1382–91. doi: 10.1017/S0033291720000173

114. McIntyre R, Filteau M, Martin L, Patry S, Carvalho A, Cha D, et al. Treatment-resistant depression: definitions, review of the evidence, and algorithmic approach. *J Affect Disord.* (2014) 156:1–7.

115. Choe C, Emslie G, Mayes T. Depression. *Child Adolesc Psychiatr Clin N Am.* (2012) 21:807–29.

116. Cipriani A, Furukawa T, Salanti G, Chaimani A, Atkinson L, Ogawa Y, et al. Comparative efficacy and acceptability of 21 antidepressant drugs for the acute treatment of adults with major depressive disorder: a systematic review and network meta-analysis. *Lancet.* (2018) 391:1357–66.

117. van Dinteren R, Arns M, Kenemans L, Jongsma M, Kessels R, Fitzgerald P, et al. Utility of event-related potentials in predicting antidepressant treatment response: an iSPOT-D report. *Eur Neuropsychopharmacol.* (2015) 25:1981–90. doi: 10.1016/j.euroneuro.2015.07.022

118. Williams L, Rush A, Koslow S, Wisniewski S, Cooper N, Nemeroff C, et al. International study to predict optimized treatment for depression (iSPOT-D), a randomized clinical trial: rationale and protocol. *Trials.* (2011) 5:12. doi: 10.1186/1745-6215-12-4

119. Chekroud A, Zotti R, Shehzad Z, Gueorguieva R, Johnson M, Trivedi M, et al. Cross-trial prediction of treatment outcome in depression: a machine learning approach. *Lancet Psychiatry.* (2016) 3:243–50.

120. Korgaonkar M, Rekshan W, Gordon E, Rush A, Williams L, Blasey C, et al. Magnetic resonance imaging measures of brain structure to predict antidepressant treatment outcome in major depressive disorder. *EBioMedicine.* (2015) 2:37–45.

121. Rifkin R, Klautau A, Rifkin R. In defense of one-vs-all classification. *J Mach Learn Res.* (2004) 5:101–41.

122. DeRubeis R, Cohen Z, Forand N, Fournier J, Gelfand L, Lorenzo-Luaces L. The personalized advantage index: translating research on prediction into individualized treatment recommendations. A demonstration. *PLoS One.* (2014) 9:e83875. doi: 10.1371/journal.pone.0083875

123. Datta A, Bansal V, Diaz J, Patel J, Reato D, Bikson M. Gyri-precise head model of transcranial direct current stimulation: improved spatial focality using a ring electrode versus conventional rectangular pad. *Brain Stimul.* (2009) 2:201–7, 7.e1. doi: 10.1016/j.brs.2009.03.005

124. Martin D, Bakir A, Lin F, Francis-Taylor R, Alduraywish A, Bai S, et al. Effects of modifying the electrode placement and pulse width on cognitive side effects with unilateral ECT: a pilot randomised controlled study with computational modelling. *Brain Stimul.* (2021) 14:1489–97. doi: 10.1016/j.brs.2021.09.014

125. Butson C, Cooper S, Henderson J, Mcintyre C. Patient-specific analysis of the volume of tissue activated during deep brain stimulation. *Neuroimage.* (2007) 34:661–70.

126. Finn E, Shen X, Scheinost D, Rosenberg M, Huang J, Chun M, et al. Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. *Nat Neurosci.* (2015) 18:1664–71.

127. Xia C, Ma Z, Ciric R, Gu S, Betzel R, Kaczkurkin A, et al. Linked dimensions of psychopathology and connectivity in functional brain networks. *Nat Commun.* (2018) 9:3003.

128. Koutsouleris N, Dwyer D, Degenhardt F, Maj C, Urquijo-Castro M, Sanfelici R, et al. Multimodal machine learning workflows for prediction of psychosis in patients with clinical high-risk syndromes and recent-onset depression. *JAMA Psychiatry.* (2021) 78:195–209. doi: 10.1001/jamapsychiatry.2020.3604

129. Borsboom D, Cramer A. Network analysis: an integrative approach to the structure of psychopathology. *Annu Rev Clin Psychol.* (2013) 9:91–121. doi: 10.1146/annurev-clinpsy-050212-185608

130. Huys Q, Guitart-Masip M, Dolan R, Dayan P. Decision-theoretic psychiatry. *Clin Psychol Sci.* (2015) 3:400–21.

131. Goldbeter A. Origin of cyclicity in bipolar disorders: a computational approach. *Pharmacopsychiatry.* (2013) 46(Suppl. 1):22–5. doi: 10.1055/s-0033-1341502

132. Berwian I, Wenzel J, Collins A, Seifritz E, Stephan K, Walter H, et al. Computational mechanisms of effort and reward decisions in patients with depression and their association with relapse after antidepressant discontinuation. *JAMA Psychiatry.* (2020) 77:513–22. doi: 10.1001/jamapsychiatry.2019.4971

133. O'reilly R, Frank M. Making working memory work: a computational model of learning in the prefrontal cortex and basal ganglia. *Neural Comput.* (2006) 18:283–328. doi: 10.1162/089976606775093909

134. Loosen A, Hauser T. Towards a computational psychiatry of juvenile obsessive-compulsive disorder. *Neurosci Biobehav Rev.* (2020) 118:631–42. doi: 10.1016/j.neubiorev.2020.07.021

135. Adams R, Pinotsis D, Tsirlis K, Unruh L, Mahajan A, Horas A, et al. Computational modeling of electroencephalography and functional magnetic resonance imaging paradigms indicates a consistent loss of pyramidal cell synaptic gain in schizophrenia. *Biol Psychiatry.* (2022) 91:202–15. doi: 10.1016/j.biopsych.2021.07.024

136. Cavanagh S, Lam N, Murray J, Hunt L, Kennerley S. A circuit mechanism for decision-making biases and NMDA receptor hypofunction. *eLife.* (2020) 9:e53664. doi: 10.7554/eLife.53664

137. Anticevic A, Lisman J. How can global alteration of excitation/inhibition balance lead to the local dysfunctions that underlie schizophrenia? *Biol Psychiatry.* (2017) 81:818–20. doi: 10.1016/j.biopsych.2016.12.006

138. Unal G, Swami J, Canela C, Cohen S, Khadka N, FallahRad M, et al. Adaptive current-flow models of ECT: explaining individual static impedance, dynamic impedance, and brain current density. *Brain Stimul.* (2021) 14:1154–68. doi: 10.1016/j.brs.2021.07.012

139. Valton V, Romaniuk L, Douglas Steele J, Lawrie S, Seriès P. Comprehensive review: computational modelling of schizophrenia. *Neurosci Biobehav Rev.* (2017) 83:631–46.

140. Csikszentmihalyi M. Reflections on the field. roeper review. *J Gift Educ.* (1998) 21:80–1.

141. Guilford J. Creativity: yesterday, today and tomorrow. *J Creat Behav.* (1967) 1:3–14.

142. Mekern V, Hommel B, Sjoerds Z. Computational models of creativity: a review of single-process and multi-process recent approaches to demystify creative cognition. *Curr Opin Behav Sci.* (2019) 27:47–54. doi: 10.1016/j.cobeha.2018.09.008

143. Gillan C, Seow T. Carving out new transdiagnostic dimensions for research in mental health. *Biol Psychiatry.* (2020) 5:932–4. doi: 10.1016/j.bpsc.2020.04.013

144. Cuthbert B, Insel T. Toward the future of psychiatric diagnosis: the seven pillars of RDoC. *BMC Med.* (2013) 11:126. doi: 10.1186/1741-7015-11-126

145. Cuthbert B. The RDoC framework: facilitating transition from ICD/DSM to dimensional approaches that integrate neuroscience and psychopathology. *World Psychiatry.* (2014) 13:28–35. doi: 10.1002/wps.20087

146. Mittal V, Wakschlag L. Research domain criteria (RDoC) grows up: strengthening neurodevelopment investigation within the RDoC framework. *J Affect Disord.* (2017) 216:30–5. doi: 10.1016/j.jad.2016.12.011

147. Wiecki T, Poland J, Frank M. Model-based cognitive neuroscience approaches to computational psychiatry: clustering and classification. *Clin Psychol Sci.* (2015) 3:378–99. doi: 10.1186/s12868-016-0283-6

148. Lass A, Jordan D, Winer E. Using theory to guide exploratory network analyses. *J Clin Psychol.* (2022) 79:531–40.

149. Michelini G, Palumbo I, DeYoung C, Latzman R, Kotov R. Linking RDoC and HiTOP: a new interface for advancing psychiatric nosology and neuroscience. *Clin Psychol Rev.* (2021) 86:102025. doi: 10.1016/j.cpr.2021.102025

150. Viani N, Kam J, Yin L, Bittar A, Dutta R, Patel R, et al. Temporal information extraction from mental health records to identify duration of untreated psychosis. *J Biomed Semant.* (2020) 11:2. doi: 10.1186/s13326-020-00220-2

151. Corvin A, Sullivan P. What next in schizophrenia genetics for the psychiatric genomics consortium? *Schizophr Bull.* (2016) 42:538–41. doi: 10.1093/schbul/sbw014

152. Haro J, Ayuso-Mateos J, Bitter I, Demote-Mainard J, Leboyer M, Lewis S. ROAMER: roadmap for mental health research in Europe JOSEP. *Int J Methods Psychiatr Res.* (2014) 23(Suppl. 1):1–14. doi: 10.1002/mpr.1406

153. Caldas de Almeida J. Mental health services development in latin america and the caribbean: achievements, barriers and facilitating factors. *Int Health.* (2013) 5:15–8. doi: 10.1093/inthealth/ihs013

154. Looijestijn J, Bloma J, Aleman A, Hoek H, Goekoop R. An integrated network model of psychotic symptoms. *Neurosci Biobehav Rev.* (2015) 59:238–50.

155. Mota N, Resende A, Mota-Rolim S, Copelli M, Ribeiro S. Psychosis and the control of lucid dreaming. *Front Psychol.* (2016) 7:294. doi: 10.3389/fpsyg.2016.00294

156. Bertola L, Mota N, Copelli M, Rivero T, Diniz B, Watt D, et al. Graph analysis of verbal fluency test discriminate between patients with Alzheimer ' s disease, mild cognitive impairment and normal elderly controls. *Front Aging Neurosci.* (2014) 6:185. doi: 10.3389/fnagi.2014.00185

157. Simon G, Johnson E, Lawrence J, Rossom R, Ahmedani B, Lynch F, et al. Predicting suicide attempts and suicide deaths following outpatient visits using electronic health records. *Am J Psychiatry.* (2018) 175:951–60.

158. Wang L, Chignell M, Zhang Y, Shan B, Sheehan K, Razak F, et al. Boosting delirium identification accuracy with sentiment based natural language processing. *JMIR Med Inform.* (2022) 10:e38161. doi: 10.2196/38161

159. Liu N, Luo K, Yuan Z, Chen Y. A transfer learning method for detecting Alzheimer's disease based on speech and natural language processing. *Front Public Health.* (2022) 10:772592. doi: 10.3389/fpubh.2022.772592

160. Yeung A, Iaboni A, Rochon E, Lavoie M, Santiago C, Yancheva M, et al. Correlating natural language processing and automated speech analysis with clinician assessment to quantify speech-language changes in mild cognitive impairment and Alzheimer's dementia. *Alzheimers Res Ther.* (2021) 13:109. doi: 10.1186/s13195-021-00848-x

161. Joyce C, Markossian T, Nikolaides J, Ramsey E, Thompson H, Rojas J, et al. The evaluation of a clinical decision support tool using natural language processing to screen hospitalized adults for unhealthy substance use: protocol for a quasi-experimental design. *JMIR Res Protoc.* (2022) 11:e42971. doi: 10.2196/42971

162. Kishimoto T, Nakamura H, Kano Y, Eguchi Y, Kitazawa M, Liang K, et al. Understanding psychiatric illness through natural language processing (UNDERPIN): rationale, design, and methodology. *Front Psychiatry.* (2022) 13:954703. doi: 10.3389/fpsyt.2022.954703

163. Khodayari-Rostamabad A, Reilly J, Hasey G, de Bruin H, MacCrimmon DJ. A machine learning approach using EEG data to predict response to SSRI treatment for major depressive disorder. *Clin Neurophysiol.* (2013) 124:1975–85. doi: 10.1016/j.clinph.2013.04.010

164. Bai S, Dokos S, Ho K, Loo C. A computational modelling study of transcranial direct current stimulation montages used in depression. *Neuroimage.* (2014) 87:332–44.

165. Peterchev A, Rosa M, Deng Z, Prudic J, Lisanby S. Electroconvulsive therapy stimulus parameters: rethinking dosage. *J ECT.* (2010) 26:159–74. doi: 10.1097/YCT.0b013e3181e48165

TYPE Original Research
PUBLISHED 09 March 2023
DOI 10.3389/fpsyt.2023.1121194

OPEN ACCESS

EDITED BY
Jianzhong Su,
University of Texas at Arlington,
United States

REVIEWED BY
Lisa Wagels,
University Hospital RWTH Aachen,
Germany
Ti-Fei Yuan,
Shanghai Jiao Tong University,
China

*CORRESPONDENCE
Lian Duan
✉ duan_lian@szu.edu.cn
Pengfei Xu
✉ pxu@bnu.edu.cn

SPECIALTY SECTION
This article was submitted to
Computational Psychiatry,
a section of the journal
Frontiers in Psychiatry

RECEIVED 11 December 2022
ACCEPTED 16 February 2023
PUBLISHED 09 March 2023

CITATION
Ai H, Duan L, Huang L, Luo Y, Aleman A and Xu P (2023) Dissociated deficits of anticipated and experienced regret in at-risk suicidal individuals.
Front. Psychiatry 14:1121194.
doi: 10.3389/fpsyt.2023.1121194

Dissociated deficits of anticipated and experienced regret in at-risk suicidal individuals

Hui Ai[1,2], Lian Duan[3]*, Lin Huang[3], Yuejia Luo[3,4,5], André Aleman[3,6] and Pengfei Xu[4,5]*

[1]Institute of Applied Psychology, Tianjin University, Tianjin, China, [2]Academy of Medical Engineering and Translational Medicine, Tianjin University, Tianjin, China, [3]Shenzhen Key Laboratory of Affective and Social Neuroscience, Center for Brain Disorders and Cognitive Sciences, Shenzhen University, Shenzhen, China, [4]Beijing Key Laboratory of Applied Experimental Psychology, National Demonstration Center for Experimental Psychology Education (BNU), Faculty of Psychology, Beijing Normal University, Beijing, China, [5]Center for Emotion and Brain, Shenzhen Institute of Neuroscience, Shenzhen, China, [6]Section Cognitive Neuroscience, Department of Biomedical Sciences of Cells and Systems, University Medical Center Groningen, University of Groningen, Groningen, Netherlands

Backgrounds: Decision-making deficits have been reported as trans-diagnostic characteristics of vulnerability to suicidal behaviors, independent of co-existing psychiatric disorders. Individuals with suicidal behaviors often regret their decision to attempt suicide and may have impairments in future-oriented processing. However, it is not clear how people with suicidal dispositions use future-oriented cognition and past experience of regret to guide decision-making. Here, we examined the processes of regret anticipation and experience in subclinical youth with and without suicidal ideation during value-based decision-making.

Methods: In total, 80 young adults with suicidal ideation and 79 healthy controls completed a computational counterfactual thinking task and self-reported measures of suicidal behaviors, depression, anxiety, impulsivity, rumination, hopelessness, and childhood maltreatment.

Results: Individuals with suicidal ideation showed a reduced ability to anticipate regret compared to healthy controls. Specifically, suicidal ideators' experience of regret/relief was significantly different from that of healthy controls upon obtained outcomes, while their disappointment/pleasure experience was not significantly different from healthy controls.

Conclusion: These findings suggest that young adults with suicidal ideation have difficulty predicting the consequences or the future value of their behavior. Individuals with suicidal ideation showed impairments in value comparison and flat affect to retrospective rewards, whereas individuals with high suicidality showed blunted affect to immediate rewards. Identifying the counterfactual decision-making characteristics of at-risk suicidal individuals may help to elucidate measurable markers of suicidal vulnerability and identify future intervention targets.

KEYWORDS

suicide, regret, counterfactual thinking, computational modeling, at-risk youths

Introduction

Suicide is the second leading cause of death among adolescents (1), and nearly one-third of suicides occur among young people (2). Heterogeneous risk factors including early-life adversity, psychopathology, and stressful life events can increase suicide risk (3). Given the multifactorial nature of its etiology, which research has yet to fully elucidate, it is difficult to predict suicidal behavior. For example, although major depressive disorder and substance use have been reported as important risk factors for suicide (4), many patients with these conditions do not exhibit suicidal behavior. The assessment of suicide risk is largely dependent on individuals' self-perception and willingness to report suicidal behaviors (5). Furthermore, adequate intervention for these psychopathologies may not prevent suicide *per se*. It has been proposed that suicidal behavior is an endophenotype that should be studied and treated independently of specific psychiatric disorders (6, 7). Identifying specific risk factors for suicidal behaviors is the first step toward early detection and prevention in at-risk individuals (8).

Decision-making alterations have been found not only in patients with suicidal behaviors related to mood disorders (7, 9–14), but also in suicidal patients with PTSD (15) and schizophrenia [reviewed by (16)]. Moreover, psychiatric patients with suicidal behaviors have shown distinctive impairments during decision-making compared to patients without suicidal behaviors (7, 17). These findings suggest that altered decision-making may be a potential trans-diagnostic marker of suicide, independent of co-existing psychopathologies (8, 18). Although studies on decision-making in suicide have mostly focused on past experiences, assessing reactions to future events may be important for the early detection of suicide (19). Previous studies have shown that suicide attempters differ from non-attempters in future-oriented cognition, characterized by overestimating negative future events and forecasting less happiness for positive future events (19, 20). However, it is unclear how this future-oriented cognition affects present decision-making in suicidal individuals, or whether they are able to use prospective outcomes to guide action selection. Identifying the future-oriented decision-making in suicidal individuals without a diagnosis of psychiatric disorder will help to uncover the cognitive mechanism of suicide independent of diagnosis.

Suicidal individuals often regret their decision to engage in suicidal behavior (21), which involves counterfactual decision-making as well as a regret response. Counterfactual thinking refers to thoughts that compare possible outcomes of alternative choices in the past with the current situation, which often occurs in goal-directed decision-making and coexists with the experience of regret (22). While people generally avoid regret by choosing the option with the least expected regret (23), it has been proposed that suicidal individuals are overly sensitive to self-blamed regret about past events, which may expose them to intense internal conflict and trigger suicidal behavior (24). However, it is unclear whether suicidal individuals have deficits in anticipating or experiencing regret for future events.

Elucidating the future-oriented decision-making and affective processing mechanisms in suicidal at-risk individuals is important for understanding the progression of suicidal behaviors and is the first step toward early detection. Therefore, the aim of our study was to investigate the regret processing in suicide by using a counterfactual thinking paradigm in suicidal at-risk youth without psychiatric disorders. This paradigm has been well-designed to observe value-based predictive behavior as well as emotional responses to counterfactual outcomes (25, 26). Given that suicidal behavior has been reported to be associated with impairments in emotion processing (27), and that healthy individuals tend to be regret-avoidant (28), we predict that individuals with suicidal ideation will show altered emotional responses to outcomes compared to non-suicidal controls. We also predict that suicidal individuals would show impaired performance during counterfactual decision-making since the expected reward value has been found to be disrupted in previous studies of suicidal behavior (10, 29).

Methods and materials

Participants

Participants completed the Scale for Suicidal Ideation [SSI-19; (30)]. The SSI-19 is a 19-item scale designed to measure suicidal ideation or intent. Current suicidal ideation in the last 2 weeks and suicidal ideation at the worst point in life were assessed. For those with suicidal ideation, an explicit question on suicide attempts was asked to assess whether they had ever attempted suicide. Because of the high comorbidity between suicidality and psychiatric conditions such as depression and anxiety, participants were instructed to complete the Beck Depression Inventory-II [BDI-II; (31)] and Spielberger's State–Trait Anxiety Inventory (32). Participants with depressive states (BDI scores above 14) were excluded. The suicidal ideation and control groups were matched for level of state anxiety. Moreover, to control for the possible confounding effects of childhood maltreatment, rumination, hopelessness, and impulsivity (5), participants completed the Childhood Trauma Questionnaire (33), the Rumination Reconsidered scales (34), the Beck Hopelessness Scale (35), and the Barratt Impulsiveness Scale (36). Participants with a diagnosis or family history of mental disorders were excluded from our study. They were also screened with the exclusion criteria of alcohol or substance use and any history of neurological illness.

Task paradigm

The current counterfactual-thinking task was adapted from those of Baskin-Sommers et al. (37), Gillan et al. (38), and Camille et al. (26). Prior to the experiment, participants were instructed to maximize their score in order to receive more rewards. On each trial, participants were asked to choose one out of two wheels (Figure 1). The proportions of different colors (0.25, 0.5, or 0.75) represented the probability of getting the particular points. There were 16 possible outcomes for each option: −210, 210; −210, 70; −210, −70; −210, −210; −70, 210; −70, 70; −70, −70; −70, −210; 70, 210; 70, 70; 70, −210; 210, 210; 210, 70; 210, −70; 210, −210. To control for between-subject differences in the presentation of trials, each participant received the same order of trials as the others and the probability of outcomes was not randomized.

To exacerbate the regret effect (26), participants had the opportunity to change their mind in 50% of the trials. Once the participant had chosen one of the two wheels, the unchosen wheel was darkened and the chosen wheel was highlighted. After the outcome was presented, a 9-point rating scale appeared on the screen asking participants to rate how they felt about the outcome of the chosen option. The aim of this rating was to assess the emotional experience of counterfactual thinking in relation to achieving another outcome within the same wheel. Following this partial feedback, the outcome of the unchosen option was presented and participants were asked to rate their feelings on a second 9-point rating scale as complete feedback. This rating was designed to measure the emotion resulting from counterfactual thinking on what would have happened if the other option wheel had been chosen. After completing all 80 trials, the participant's final score was presented on the screen. We used Psychotoolbox-3[1] to present the stimuli and record the behavioral responses.

Data analysis

Emotion rating scores

For the first rating on partial feedback, we calculated the obtained outcome and the difference between the obtained and unobtained outcomes in the same wheel ({obtained outcome > unobtained outcome of the same wheel} was operationalized as chance counterfactual, indicating the differences between the obtained value and what the participant could have obtained within the chosen wheel). For the second rating on complete feedback, we calculated the obtained outcome and the difference between the obtained and unobtained outcomes in the other wheel ({obtained > unobtained outcome in the other option} was operationalized as agent counterfactual, indicating the differences between the obtained value and what the participant could have obtained if chosen the other wheel). After this, we built linear mixed effect models with lme4 package in R (version 3.6.2) for two rating outcomes with the groups (suicide, control), outcomes of each trail, and chance counterfactuals as fixed-effect predictors, group × chance counterfactual and group × obtained outcome as interaction terms, and participants as a random factor in rating model 1; with group, outcome of each trial, agent counterfactuals as fixed-effect predictors, participants as a random factor, and group × obtained outcome, group × agent counterfactuals as interaction terms in rating model 2.

Option-selection modeling

We modeled the counterfactual behaviors by estimating the following three factors guiding decision-making: expected value (EV), expected disappointment (ED), and expectation of regret/relief (regret/relief, R). Here, x_1 and y_1 represent two possible outcomes of option 1 and $x_1 > y_1$; x_2 and y_2 represent two possible outcomes of option 2, and $x_2 > y_2$; p and 1-p represent the possibilities of obtaining x_1 and y_1; q and 1-q represent the probability to get x_2 and y_2 in option 2.

With these parameters, we first calculated the maximal expected value with Eq. (1), where $EV > 0$ indicates a higher EV in option 1 than in option 2.

$$\begin{aligned}\mathrm{EV} = \mathrm{EV}_{o1} - \mathrm{EV}_{o2} &= \left[p * x_1 + (1-p) * y_1\right] \\ &-\left[\mathrm{q} * x_2 + (1-q) * y_2\right]\end{aligned} \quad (1)$$

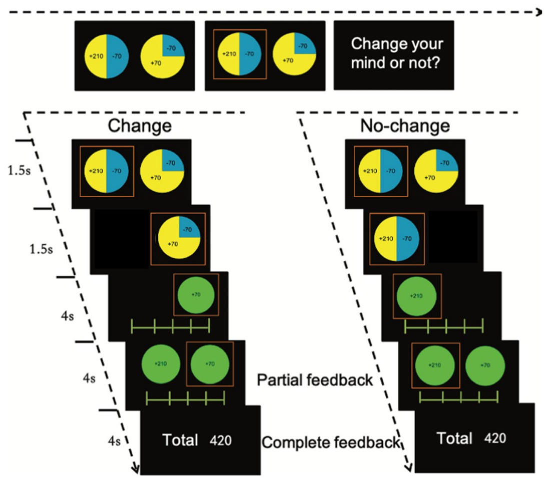


FIGURE 1
Task procedure. The proportion of two colors in one panel represents the probability of two outcomes. Participants were asked to make a choice and have a 50% opportunity to change their mind. Then, they were asked to rate how they feel after they make the choice (first rating) and after the presentation of the outcomes of the other panel (second rating).

We then calculated the expected disappointment (ED) of each trial with Eq. (2), where ED_{o1} and ED_{o2} represent the estimate of expected disappointment for option 1 and option 2, and $EDo_2 > EDo_1$ indicates the participant should choose option 1 when trying to avoid future disappointment.

$$\mathrm{ED} = \mathrm{ED}_{o2} - \mathrm{ED}_{o1} = (x_2 - y_2)(1-q) - (x_1 - y_1)(1-p) \quad (2)$$

Next, we calculated the difference between the possible highest and lowest outcomes of the two options as the index of expected regret/relief. This calculation was based on the assumption that the difference between the obtained outcome and the possible outcome if one chose differently would cause the participant's regret or relief. The bigger the difference, the more intense the regret or relief. $R > 0$ indicates lower regret/relief from option 1:

$$\mathrm{R} = (y_1 - x_2) - (y_2 - x_1) \quad (3)$$

The probability of choosing option 1 for each trial of each participant (t, trail number; i, participant number) was calculated as:

$$\mathrm{P}(\mathrm{O}_{1ti}) = 1 - \mathrm{P}(\mathrm{O}_{2ti}) = \mathrm{F}(\mathrm{ED}_{ti}, \mathrm{R}_{ti}, \mathrm{EV}_{ti}) \quad (4)$$

F denotes the inverse logit function to estimate individual expected value, risk variance, and regret. The probability of choosing option 2 was modeled in the same way. We used a linear mixed effect (LME) logistic regression model in R, with EV, ED, and R as continuous fixed-effect factors, the group as a fixed-effect factor, the participant as a

[1] http://psychtoolbox.org/

TABLE 1 Demographics description of participants.

Group	ISD	HC	*t*	Chi-square	*P* value
Sample size (*N*)	80	79	–	–	–
Age *Mean (SD)*	19.96 (1.36)	20.14 (1.52)	−0.77	–	0.44
Sex (male/female) (*N*)	35/45	39/40	–	0.50	0.53
Scale for Suicide Ideation_worst *Mean (SD)*	14.24 (7.09)	2.37 (4.48)	12.61	–	<0.05*
Scale for Suicide Ideation_current *Mean (SD)*	2.29 (3.84)	0.67 (1.80)	3.29	–	0.05*
Suicide attempts *(yes/no)*	16/64	0/0	–	–	-
BDI *Mean (SD)*	9.93 (7.08)	8.20 (6.32)	1.62	–	0.11
S-AI *Mean (SD)*	40.43 (10.12)	39.91 (9.35)	0.33	–	0.74
T-AI *Mean (SD)*	44.64 (9.29)	43.47 (8.34)	0.84	–	0.41
BIS *Mean (SD)*	59.83 (8.93)	60.00 (7.49)	−0.13	–	0.89
BHS *Mean (SD)*	5.81 (3.04)	5.67 (3.64)	0.23	–	0.79
CTQ *Mean (SD)*	40.63 (11.44)	39.04 (11.11)	0.89	–	0.38
RRS *Mean (SD)*	47.60 (9.73)	45.75 (7.07)	1.38	–	0.17

ISD, individuals with suicidal dispositions; HC, healthy controls; BDI, Beck Depression Inventory; RRS, Rumination Reconsidered scales; S-AI, Spielberger's State anxiety inventory; T-AI, Spielberger's State anxiety inventory; BIS, Barratt impulsiveness scale; CTQ, Childhood Trauma Questionnaire. *$p < 0.05$.

random-effect factor, and choice as the binary outcome variable. Another LME logistic regression model was built to test the main effects and interactions among three estimated parameters and groups. Besides the full model, we built multiple models by reducing factors stepwise to check the factor contribution. Likelihood ratio tests were used to confirm statistical significance when comparing models with and without terms of interest. The results were regarded as significant at *$p < 0.05$*. The criteria to find the best model was the AIC (Akaike Information Criterion) value for each model (39).

To further clarify interactions of suicidal severity with three estimated parameters, we did a sensitivity analysis by building a model with SSI scores at the worst point and EV, ED, and R as continuous fix-effect factors, participant as a random-effect factor, and choice as the outcome variable.

To control for the possible effect of a depressive state, anxiety state and trait, impulsivity, hopelessness, rumination, and childhood maltreatment on the task, we set them as covariates. To check the collinearity of our task parameters, we tested correlations between the slopes of the task parameters and these covariates. Furthermore, to test the effect of change-of-mind, we calculated the frequency of change and repeated the analysis on the rating from the complete feedback by including binary factor (change or not change) in the model as an interaction term.

Results

Sample characteristics

There were 202 participants who completed the questionnaires. To match the depressive and anxious levels, rumination, hopelessness, impulsivity, and experience of childhood trauma (CTQ) between the two groups, we excluded 36 healthy controls and 5 suicidal ideators with BDI scores above 14. In total, 80 participants who reported having suicidal ideation at their worst point in life (45 females, age 19.96 ± 1.36) were grouped as individuals with a suicidal disposition (ISD). The mean suicidal ideation score was 14.24 ($SD = 7.09$). Furthermore, 16 participants had past suicidal attempts (8 females, age 20.56 ± 1.03). Suicide attempters were different from suicidal ideators in suicidal intention scores ($t = 5.53$, $p < 0.05$). Finally, 79 individuals without any suicidal disposition or psychiatric problems were grouped as controls (HC) (Demographics, Table 1).

Disappointment affect ratings

There was a significant main effect of the group on affective responses to partial feedback. Additionally, a main effect of the obtained outcome on affective responses to partial feedback was also found, with a low obtained outcome related to negative affect and a high obtained outcome related to positive affect in both groups (Figure 2A). We also found a main effect of chance counterfactual on affective responses to partial feedback, with a larger obtained outcome than counterfactual outcome associated with a more positive affect, and a lower obtained outcome than counterfactual outcome associated with a more negative affect (Figure 2B). There was no interaction between the group and obtained outcomes or interaction between the group and chance obtained outcomes to partial feedback (Table 2).

To examine the effect of suicidal severity on affect ratings to partial feedback, we did a sensitivity analysis by setting scores of Scales for Suicidal Ideation (SSI) as a continuous fix-effect predictor instead of the group with participants with suicidal dispositions. A significant main effect for the obtained outcome ($\text{Beta} = 1.23 \times 10^{-2}$, $SE = 5.95 \times 10^{-4}$, $95\%CI = 1.11 \times 10^{-2}$ to 0.0135, $t = 20.71$, $p < 0.001$) as well as a significant interaction between the obtained outcome and the scores for suicidal ideation ($\text{Beta} = -1.15 \times 10^{-4}$, $SE = 3.71 \times 10^{-5}$, $95\%CI = -1.88 \times 10^{-4}$ to -4.26×10^{-5}, $t = -3.11$, $p = 0.002$) were found (Figure 3A). A significant main effect for chance counterfactuals was also found ($\text{Beta} = 2.049 \times 10^{-2}$, $SE = 3.254 \times 10^{-4}$, $95\%CI = 0.01$ to

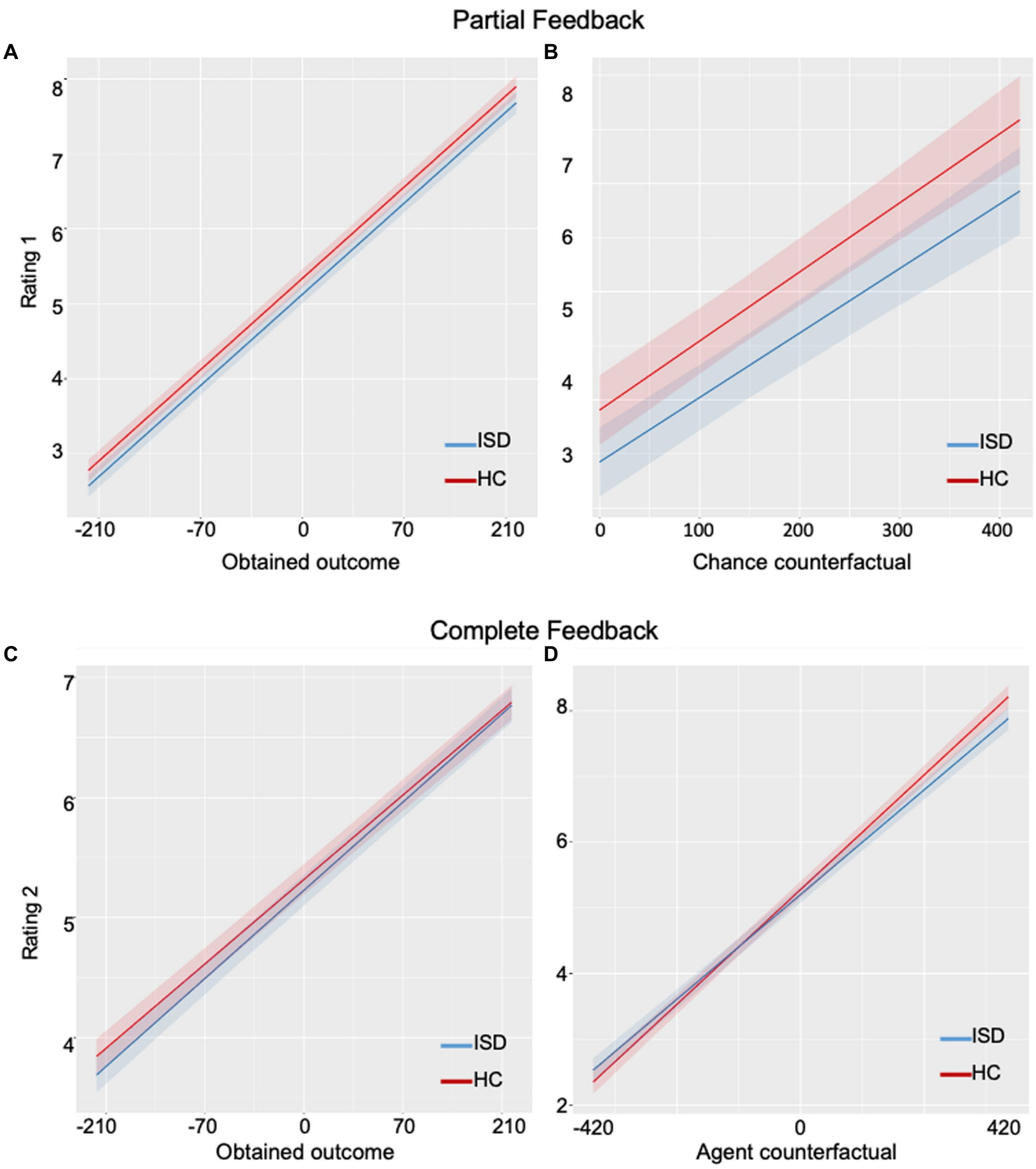


FIGURE 2
Plots of affective ratings on partial and complete feedback in the suicidal group and the healthy group. **(A)** The rate of disappointment/pleasure upon obtained outcome was not significantly different between groups. **(B)** The rate of disappointment/pleasure upon chance counterfactual outcome was not significantly different between the two groups. **(C)** The rate of regret/relief upon obtained outcome was not significantly different between groups. **(D)** The rate of regret/relief upon agent counterfactual outcome was significantly different between groups: the suicidal group showed blunted responses compared to the healthy group. ISD, individuals with suicidal dispositions; HC, healthy controls.

0.03, $t = 6.30$, $p < 0.001$). No main effect for suicidal scores was found (Beta = 1.076×10^{-2}, SE = 8.36×10^{-3}, 95%CI = -5.62×10^{-3} to 2.71×10^{-2}, $t = 1.29$, $p = 0.20$) nor interaction between chance counterfacutals and suicidal ideations (Beta = -3.84×10^{-6}, SE = 2.03×10^{-5}, 95%CI = -4.36×10^{-5} to 3.60×10^{-5}, $t = -0.19$, $p = 0.85$; Figure 3B).

Overall, for partial feedback, the ISD group showed blunted emotions compared to HC participants (Figures 2A,B). Within the ISD group, we found that individuals with higher suicidal ideation scores (high-suicidality) had a more blunted affect to the obtained outcomes than individuals with lower suicidal ideations (low-suicidality; Figure 3A).

Regret affect ratings

We observed significant main effects of obtained outcome and agent counterfactual on affect rating to complete feedback across all groups (Figure 2C), with a low obtained outcome associated with a stronger negative affect and a high obtained outcome associated with a stronger positive affect. A larger obtained than unobtained outcome was associated with a stronger positive affect and a lesser obtained than unobtained outcome was associated with a stronger negative affect. There was also a significant interaction between the group and agent counterfactuals (Figure 2D). No effect of the group or interaction between the group and obtained outcome were observed (Table 2).

TABLE 2 Affect rating model with obtained and counterfactual outcome parameters.

Parameter	Coefficient	Standard error	95%CI	*t*	*p*
Affect rating1 model with all subjects					
Intercept	5.01	6.49×10^{-2}	4.88 to 5.14	77.24	<0.05*
Obtained outcome	1.22×10^{-2}	1.83×10^{-4}	1.18×10^{-2} to 1.25×10^{-2}	66.48	<0.05*
Chance counterfactuals	2.56×10^{-3}	1.91×10^{-4}	2.18×10^{-3} to 2.93×10^{-3}	13.37	<0.05*
Group	-1.91×10^{-1}	9.16×10^{-2}	-3.71×10^{-2} to -1.13×10^{-2}	−2.09	0.04*
Obtained outcome:group	-2.28×10^{-5}	2.60×10^{-4}	-5.33×10^{-4} to 4.87×10^{-2}	−0.09	0.93
Chance counterfactuals:group	-1.70×10^{-4}	2.71×10^{-4}	-6.70×10^{-4} to -3.61×10^{-4}	−0.63	0.53
Affect rating2 model with all subjects					
Intercept	5.07	6.45×10^{-2}	5.07 to 5.32	80.57	<0.05*
Obtained outcome	7.03×10^{-3}	1.70×10^{-3}	6.70×10^{-3} to 7.36×10^{-3}	41.44	<0.05*
Agent counterfactuals	6.98×10^{-3}	1.52×10^{-4}	6.68×10^{-3} to 7.36×10^{-3}	45.86	<0.05*
Group	-7.91×10^{-2}	9.09×10^{-2}	-2.57×10^{-1} to -9.90×10^{-2}	−0.87	0.39
Obtained outcome:group	3.09×10^{-4}	2.39×10^{-4}	-1.60×10^{-4} to 7.78×10^{-4}	1.29	0.20
Agent counterfactuals:group	-6.08×10^{-4}	2.15×10^{-4}	-1.03×10^{-3} to -1.88×10^{-4}	−2.83	<0.05*

*$p < 0.05$.

For the sensitivity analysis within the suicidal group, we set the model by adding SSI as a continuous fix-effect factor. Main effects of obtained outcome (Beta = 7.21×10^{-3}, SE = 3.87×10^{-4}, 95%CI = 6.45×10^{-3} to 7.97×10^{-3}, $t = 18.65$, $p < 0.001$) and agent counterfactual (Beta = 7.33×10^{-3}, SE = 3.47×10^{-4}, $t = 21.12$, $p < 0.001$) were found. A significant interaction between agent counterfactuals and SSI scores was also observed (Beta = -6.68×10^{-5}, SE = 2.20×10^{-5}, 95%CI = -1.09×10^{-4} to -2.37×10^{-5}, $t = -3.04$, $p = 0.002$; Figure 3D). No effect of SSI scores (Beta = 6.81×10^{-3}, SE = 9.24×10^{-3}, 95%CI = -1.13×10^{-2} to 2.49×10^{-2}, $t = 0.74$, $p = 0.46$) or interaction between SSI and obtained outcome (Beta = 8.66×10^{-6}, SE = 2.44×10^{-5}, 95%CI = -3.88×10^{-3} to 5.69×10^{-5}, $t = 0.36$, $p = 0.72$) were found (Figure 3C).

In summary, for complete feedback, the ISD group showed less pleasure than HC when the obtained outcome was larger than the unobtained outcome on the other wheel, and less regret when the obtained outcome was less than the unobtained outcome on the other wheel (Figure 2D). Within the ISD group, high-suicidality individuals showed a more blunted affect to agent counterfactuals than low-suicidality individuals (Figure 3D).

Decision-making

The Effects of the full choice model and the best choice model with computational parameters are summarized in Table 3. We observed a significant main effect of expected value (EV) and a significant main effect of avoidance of disappointment (ED). There was also a significant main effect of regret prediction (R), with participants choosing options to minimize future regret. Importantly, there was a significant interaction between the group and regret prediction. Specifically, the suicidal group showed a blunted sensitivity to future regret compared to healthy controls (Figure 4). There was no interaction between the group and the expected value or avoidance of disappointment parameters.

To examine the effect of suicide severity, we built another linear mixed-effect model with SSI scores as a continuous fixed factor within the ISD group. We found a significant effect for EV (Beta = 0.03, SE = 0.01, z = 9.56, $p < 0.001$), ED (Beta = -4.39×10^{-3}, SE = 8.38×10^{-4}, z = −5.23, $p < 0.001$) and R (Beta = 0.01, SE = 5.07×10^{-4}, z = 4.22, $p < 0.001$) in suicidal individuals. There was also a marginally significant interaction between SSI score and anticipation of future regret (R; Beta = -8.89×10^{-5}, SE = 3.18×10^{-5}, z = 2.80, $p = 0.05$), indicating that individuals with a high suicidal disposition are less sensitive to future regret than those with a low suicidal disposition.

In the correlation between task parameters and covariates, we observed a correlation between chance counterfactuals and impulsivity scores, further analysis with BIS as a covariate factor in the linear mixed effect model did not change the result of interaction between chance counterfactual and group (Beta = 6.02×10^{-5}, SE = 1.14×10^{-5}, 95%CI = -7.69×10^{-5} to -3.35×10^{-5}, $t = -2.14$, $p = 0.03$). Although significant correlations can be found within covariates or within task variables, no significant correlations were found between task parameters and other covariates (Table 4). Because the three task parameters, ED, EV, and R, were inter-correlated with each other, we checked the variance inflation factors (VIF) in the LME model. The VIFs were all below the commonly suggested cut-off

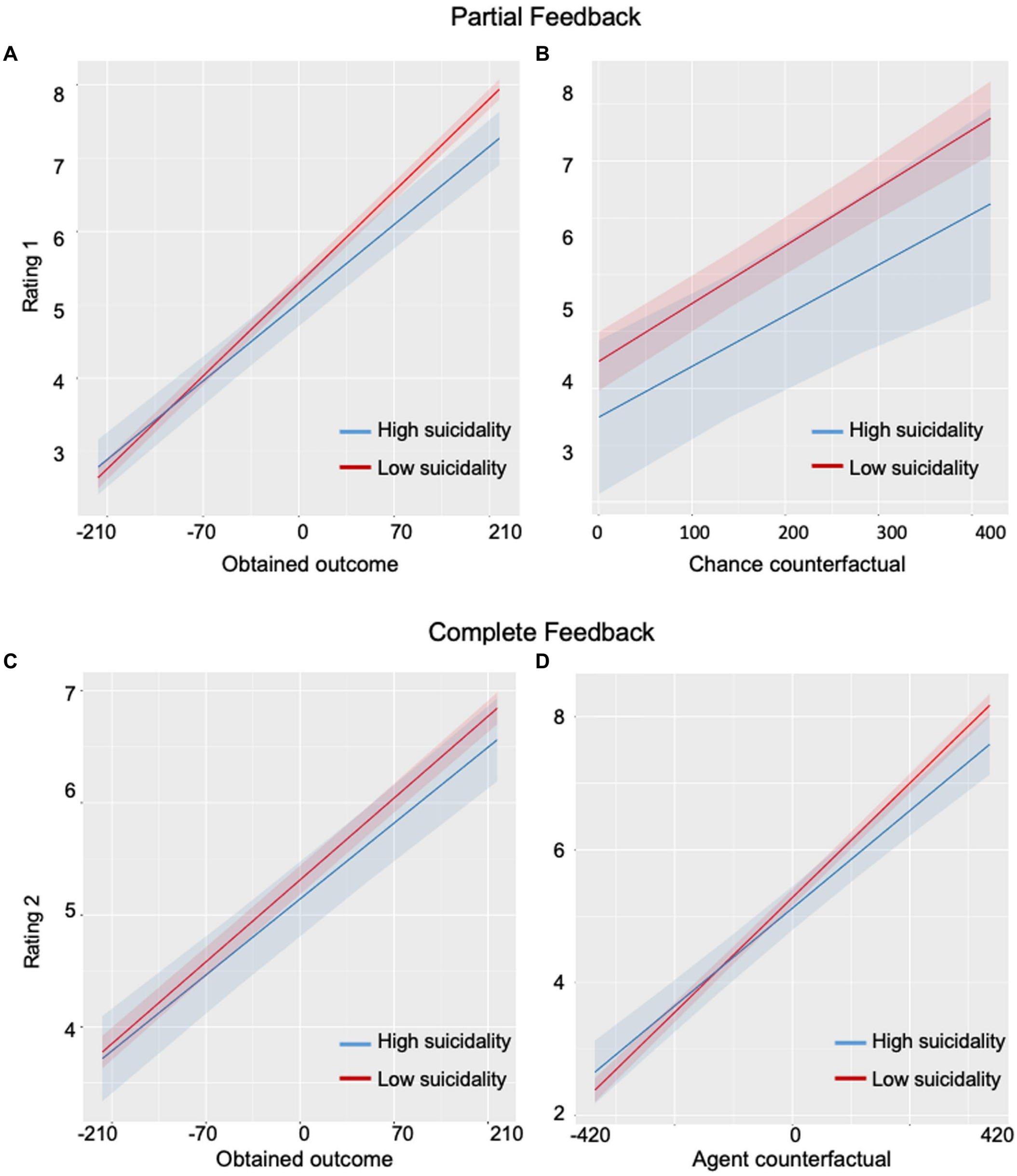


FIGURE 3
Plots of correlations between levels of suicidal ideation and emotional responses. **(A)** A significant correlation was observed between levels of suicidal ideation and rate of disappointment/pleasure upon obtained outcome: individuals with a high level of suicidal ideation showed less affect than low-suicidality individuals. **(B)** No correlation was found between levels of suicidal ideation and rate of disappointment/pleasure upon chance counterfactual outcome. **(C)** No correlation was found between levels of suicidal ideation and rate of regret/relief upon obtained outcome. **(D)** A significant correlation was found between levels of suicidal ideation and rate of regret/relief upon agent counterfactual outcome.

of 10 (VIF values were smaller when the variables were stratified), indicating that collinearity was not a problem in our model (EV:2.85, ED:4.17, R:4.18). The change-of-mind setting did not exacerbate emotional responses to the obtained outcomes ($p = 0.54$) or agent counterfactual ($p = 0.27$). Moreover, although participants had the opportunity to change their minds, very few of them did so, and even then quite infrequently (ISD group: mean = 1.87, SD = 2.73; HC group: mean = 1.94, SD = 2.28). No difference between the groups in the switching wheel rate was found ($t = 0.16$, $p = 0.87$).

Discussion

In the present study, we examined the association of suicidal ideation with regret anticipation and counterfactual emotional experience in value-based decision-making using model-based mathematical computations. Our results revealed that young adults with suicidal ideation showed a blunted anticipation of potential future regret when making decisions. Whereas suicidal ideators and past suicidal attempters showed less avoidance of future regret, young adults with more suicidal dispositions showed blunted emotional

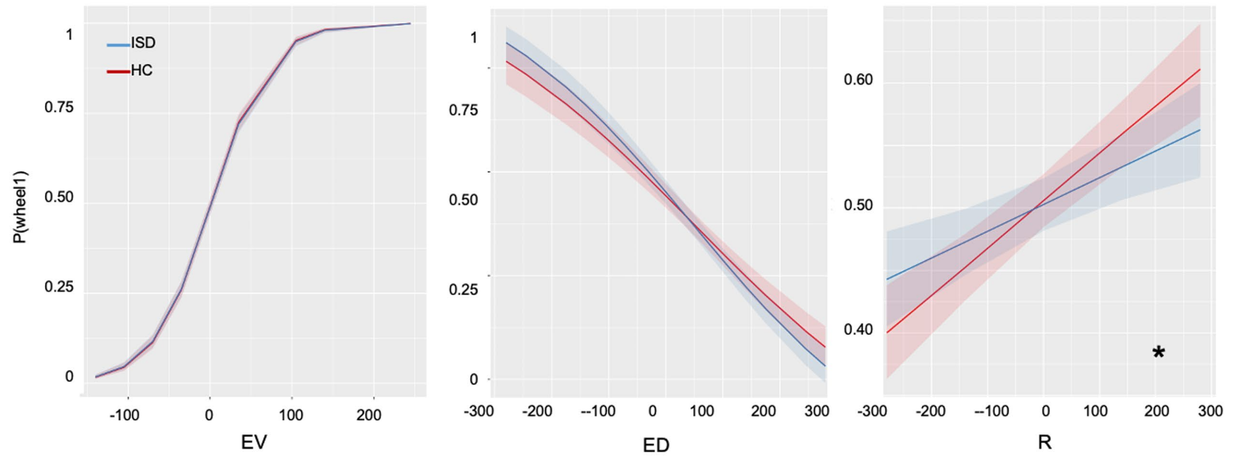


FIGURE 4
Plots on results of decision-making variables. EV, expected values; ED, expected disappointment; R, avoidance of regret. * indicating p<0.05.

TABLE 3 Choice models with computational parameters.

Parameter	Coefficient	Standard Error	95%CI	t	p
(A) The best choice model with all subjects: choice ~ E + D + R + group:D + group:R + (1 | subject)					
Intercept	5.16×10^{-1}	5.80×10^{-3}	0.50 to 0.53	88.93	<0.05*
EV	3.56×10^{-3}	8.83×10^{-5}	3.38×10^{-3} to 3.72×10^{-3}	40.25	<0.05*
ED	-6.06×10^{-4}	5.99×10^{-5}	-7.23×10^{-4} to -4.88×10^{-4}	−10.11	<0.05*
R	5.24×10^{-4}	4.02×10^{-5}	4.45×10^{-4} to 6.03×10^{-4}	13.02	<0.05*
ED:group	-1.28×10^{-4}	7.67×10^{-5}	-2.78×10^{-4} to 2.23×10^{-5}	−1.67	0.10
R:group	-1.19×10^{-4}	5.12×10^{-5}	-2.20×10^{-4} to -1.90×10^{-5}	−2.33	0.02*
159 subjects, 12,720 observations					
(B) Full choice model with all subjects: choice ~ E + D + R + group:E + group:D + group:R + (1 | subject)					
Intercept	5.16×10^{-1}	5.80×10^{-3}	0.50 to 0.53	88.91	<0.05*
EV	3.51×10^{-3}	1.25×10^{-4}	3.26×10^{-3} to 3.75×10^{-3}	28.11	<0.05*
ED	-5.95×10^{-4}	6.33×10^{-5}	-7.19×10^{-4} to -4.71×10^{-4}	−9.40	<0.05*
R	5.33×10^{-4}	4.34×10^{-5}	4.48×10^{-4} to 6.18×10^{-4}	12.27	0.59
ED:group	-1.50×10^{-4}	8.68×10^{-5}	-3.20×10^{-4} to 2.22×10^{-5}	−1.74	0.08
R:group	-1.37×10^{-4}	6.06×10^{-5}	-2.56×10^{-4} to -1.80×10^{-5}	−2.26	0.02*
159 subjects, 12,720 observations					

EV, expected values; ED, expected disappointment; R, avoidance of regret. *$p < 0.05$.

responses to the immediate outcome, regardless of win or loss (Table 4). They were also less sensitive to regret and relief in retrospective comparisons compared to healthy individuals. These results were independent of the state of depression or anxiety, the experience of childhood trauma, ruminations, hopelessness, and impulsivity, which are common risk factors for suicidality and may influence decision-making. Taken together, these findings suggest that subclinical individuals with suicidal dispositions may have specific

TABLE 4 Correlations among task parameters and covariates.

	Chance *CF*	Agent *CF*	ED	EV	R	BDI	RRS	SAI	TAI	BIS	CTQ
Chance *CF*	1	0.00*	0.79	0.25	0.33	0.99	0.34	0.69	0.77	0.01*	0.71
Agent *CF*		1	0.61	0.19	0.23	0.23	0.21	0.74	0.86	0.20	0.86
ED			1	0.00*	0.00*	0.27	0.40	0.82	0.51	0.86	0.21
EV				1	0.03*	0.88	0.57	0.33	0.77	0.12	0.47
R					1	0.43	0.42	0.65	0.75	0.42	0.57
BDI						1	0.00*	0.00*	0.00*	0.02*	0.35
RRS							1	0.00*	0.00*	0.11	0.18
SAI								1	0.00*	0.00*	0.02*
TAI									1	0.00*	0.00*
BIS										1	0.05*
CTQ											1

CF, counterfactuals; BDI, Beck Depression Inventory; RRS, Rumination Reconsidered scales; STAI, Spielberger's State–Trait anxiety inventory; BIS, Barratt impulsiveness scale; CTQ, Childhood Trauma Questionnaire; EV, expected values; ED, expected disappointment; R, avoidance of regret. *$p < 0.05$.

alterations in the use of forward prospective cognition in action-outcome comparisons to guide goal-directed behaviors and blunted emotional responses to retrospective regret cues.

In healthy individuals, anticipated regret has been suggested to guide decisions that protect one from painful consequences (23). an early clinical study has reported that psychiatric patients with suicidal behaviors have difficulty predicting the consequences or the future value of their behaviors (40) and have then been shown to have deficits in future orientation (41, 42). It has been shown that patients with lesions in the vmPFC have impairments in predicting negative outcomes, learning from negative experiences (43), and avoiding future regret (44). Moreover, dysfunctional value representation in the vmPFC has been observed in suicidal individuals, as has disrupted vmPFC-frontoparietal connectivity in reinforcement learning (45). Taken together, these findings suggest that vmPFC dysfunction might be associated with deficits in regret anticipation in suicidal individuals, including less avoidance of future regret and less consideration of negative consequences, facilitating suicidal behaviors. Our findings in subclinical suicidal ideators confirm that this disrupted anticipation of future-oriented regret might be associated with the severity of suicidality and is independent of co-existing psychiatric disorders.

Individuals with suicidal ideations showed less pleasure in winning and less disappointment in losing, retrospectively, compared to healthy individuals. This may indicate altered value comparisons and amotivated responses to retrospective outcomes. Previous research has found that more than half of suicidal attempters regret their suicidal actions (21). More importantly, the presence of subsequent counterfactual thinking (i.e., wishing that they had died *via* the suicidal acts) is predictive of eventual suicide (46). Although suicidal ideators did not show reduced emotional responses to immediate outcomes, analysis within the ISD group indicated decreased responses to immediate outcomes in individuals with high suicidal severity. It has been shown that suicidal individuals tend to selectively neglect decision-relevant value information in reward learning (29). Impaired value comparison in suicidal individuals has also been found in gambling and reinforcement learning (9, 10, 47, 48). Deficits in consummatory pleasure have been associated with suicide risk (49). Loss of interest and pleasure has been reported to be predictive of suicidal ideation independently of depression in both patients (50) and college students (51, 52). Extending previous findings, our findings of blunted experience of pleasure/relief with positive consequences as well as blunted experience of disappointment/regret with negative consequences in suicidal youths suggest that suicidal disposition might be associated with loss of motivation and flat emotion. These amotivational abnormalities may contribute to an altered value comparison between suicidal behavior and its alterations during a crisis and may potentially increase the likelihood of suicidal behavior.

Suicidal youths did not show a disturbed expected value or altered avoidance of disappointment compared to healthy youths. This might be because we controlled for the level of hopelessness between groups, which is associated with value comparison and despair. However, it has been proposed that negative future expectations, lack of general motivation, and impaired attribution of meanings to personal experiences are key components of hopelessness, which is a strong predictor of suicidal behaviors (53). People with negative expectations about the future and loss of motivation have been reported to have dysfunctions in striatal dopamine pathways, which may affect suicidal ideation (54, 55). Moreover, recent studies have also reported that the absence of positive expectations about the future rather than the global construct of hopelessness, plays a key role in suicidality (56). Therefore, our findings on blunted disappointment in the face of poorer current outcomes, as well as an intact ability to avoid future disappointment may alternatively suggest that this dissociation plays a key role in suicidal disposition.

There are limitations to our study that need to be taken into account in the interpretation of the results. First, given that our participants were recruited from university and that the subclinical suicidal group only shows lifetime suicidal behaviors, the generalization of our results needs to be cautious and our findings need to be replicated in samples of different ages and in psychiatric patients with suicidal dispositions. Second, although we have excluded participants with any diagnosis of mental disorders by an explicit question, a formal diagnosis will be preferred in a future study to control for the undiagnostic risks. Finally, given that both the experience of regret and the prediction of regret might be associated with key areas such as the orbital-frontal cortex and ventromedial prefrontal cortex, future neuroimaging studies are needed to examine neural differences underlying regret processing in suicide.

To conclude, our results suggest that a flat experience and blunted regret prediction are important characteristics in subclinical young adults with lifetime suicidal ideations. These model-based distinctive abnormalities of disappointment experience, regret experience, and regret prediction may shed light on putative trans-diagnostic mechanisms in the early stages of suicidality and may be of help to identify measurable markers of suicidal vulnerability and future intervention targets.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving human participants were reviewed and approved by Ethics committee of Center for Brain Disorders and Cognitive Sciences, Shenzhen University. The patients/participants provided their written informed consent to participate in this study.

Author contributions

PX, LD, AA, and YL contributed to the conception and design of the study. HA and LH performed the experiment. HA and LD performed the statistical analyses. HA and PX wrote the draft of the manuscript. All authors contributed to the manuscript revision, and read and approved the submitted version.

Funding

This study was supported by National Natural Science Foundation of China (Nos. 31700959, 31871137, and 31920103009), Young Elite Scientists Sponsorship Program by China Association for Science and Technology, Zhujiang Talent Project for Postdoctoral researchers, Guangdong International Cooperation and Exchanges (No. 2019A050510048), Guangdong Young Innovative Talent Project (No. 2016KQNCX149), Shenzhen-Hong Kong Institute of Brain Science-Shenzhen Fundamental Research Institutions (No. 2023SHIBS0003), Shenzhen Science and Technology Research Funding Program (Nos. JCYJ20210324093208021, JCYJ20180507183500566, JCYJ20150729104249783, and JCYJ20170412164413575), and Sanming Project of Medicine in Shenzhen (SZSM202003006).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. WHO. *Suicide in the world: Global Health estimates*. World Health Organization. (2019). Geneva. Available online at: http://www.who.int/

2. WHO. *Depression and other common mental disorders: global health estimates*. (2017). Geneva: World Health Organization, Available online at: http://www.who.int/

3. Franklin, JC, Ribeiro, JD, Fox, KR, Bentley, KH, Kleiman, EM, Huang, X, et al. Risk factors for suicidal thoughts and behaviors: a meta-analysis of 50 years of research. *Psychol Bull*. (2017) 143:187–232. doi: 10.1037/bul0000084

4. Nock, M. K., Hwang, I., Sampson, N., Kessler, R. C., Angermeyer, M., and Beautrais, A., . . . de Girolamo, G. Cross-national analysis of the associations among mental disorders and suicidal behavior: findings from the WHO world mental health surveys. *PLoS Med*, (2009). 6,:e1000123, doi: 10.1371/journal.pmed.1000123.

5. Turecki, G, Brent, DA, Gunnell, D, O'Connor, RC, Oquendo, MA, Pirkis, J, et al. Suicide and suicide risk. *Nat Rev Dis Primers*. (2019) 5:74. doi: 10.1038/s41572-019-0121-0

6. Aleman, A, and Denys, D. Mental health: a road map for suicide research and prevention. *Nature*. (2014) 509:421–3. doi: 10.1038/509421a

7. Richard-Devantoy, S, Olie, E, Guillaume, S, and Courtet, P. Decision-making in unipolar or bipolar suicide attempters. *J Affect Disord*. (2016) 190:128–36. doi: 10.1016/j.jad.2015.10.001

8. Courtet, P, Gottesman, II, Jollant, F, and Gould, TD. The neuroscience of suicidal behaviors: what can we expect from endophenotype strategies? *Transl Psychiatry*. (2011) 1:e7. doi: 10.1038/tp.2011.6

9. Clark, L, Dombrovski, AY, Siegle, GJ, Butters, MA, Shollenberger, CL, Sahakian, BJ, et al. Impairment in risk-sensitive decision-making in older suicide attempters with depression. *Psychol Aging*. (2011) 26:321–30. doi: 10.1037/a0021646

10. Jollant, F., Bellivier, F., Leboyer, M., Astruc, B., Torres, S., and Verdier, R., . . . Courtet, P. Impaired decision making in suicide attempters. *Am J Psychiatry*, (2005). 162: 304–310. doi:10.1176/appi.ajp.162.2.304

11. Jollant, F, Guillaume, S, Jaussent, I, Castelnau, D, Malafosse, A, and Courtet, P. Impaired decision-making in suicide attempters may increase the risk of problems in affective relationships. *J Affect Disord*. (2007) 99:59–62. doi: 10.1016/j.jad.2006.07.022

12. Jollant, F, Lawrence, NS, Olie, E, O'Daly, O, Malafosse, A, Courtet, P, et al. Decreased activation of lateral orbitofrontal cortex during risky choices under uncertainty is associated with disadvantageous decision-making and suicidal behavior. *NeuroImage*. (2010) 51:1275–81. doi: 10.1016/j.neuroimage.2010.03.027

13. Malloy-Diniz, LF, Neves, FS, Abrantes, SS, Fuentes, D, and Correa, H. Suicide behavior and neuropsychological assessment of type I bipolar patients. *J Affect Disord*. (2009) 112:231–6. doi: 10.1016/j.jad.2008.03.019

14. Martino, DJ, Strejilevich, SA, Torralva, T, and Manes, F. Decision making in euthymic bipolar I and bipolar II disorders. *Psychol Med*. (2011) 41:1319–27. doi: 10.1017/S0033291710001832

15. Barredo, J, Aiken, E, van't Wout-Frank, M, Greenberg, BD, Carpenter, LL, and Philip, NS. Neuroimaging correlates of suicidality in decision-making circuits in posttraumatic stress disorder. *Front Psych*. (2019) 10:44. doi: 10.3389/fpsyt.2019.00044

16. Richard-Devantoy, S, Orsat, M, Dumais, A, Turecki, G, and Jollant, F. Neurocognitive vulnerability: suicidal and homicidal behaviours in patients with schizophrenia. *Can J Psychiatr*. (2014) 59:18–25. doi: 10.1177/070674371405900105

17. Gorlyn, M, Keilp, JG, Oquendo, MA, Burke, AK, and John Mann, J. Iowa gambling task performance in currently depressed suicide attempters. *Psychiatry Res*. (2013) 207:150–7. doi: 10.1016/j.psychres.2013.01.030

18. Dombrovski, AY, and Hallquist, MN. The decision neuroscience perspective on suicidal behavior: evidence and hypotheses. *Curr Opin Psychiatry*. (2017) 30:7–14. doi: 10.1097/YCO.0000000000000297

19. Marroquin, B, Nolen-Hoeksema, S, and Miranda, R. Escaping the future: affective forecasting in escapist fantasy and attempted suicide. *J Soc Clin Psychol*. (2013) 32:446–63. doi: 10.1521/jscp.2013.32.4.446

20. Golub, SA, Gilbert, DT, and Wilson, TD. Anticipating one's troubles: the costs and benefits of negative expectations. *Emotion*. (2009) 9:277–81. doi: 10.1037/a0014716

21. Henriques, G, Wenzel, A, Brown, GK, and Beck, AT. Suicide attempters' reaction to survival as a risk factor for eventual suicide. *Am J Psychiatry*. (2005) 162:2180–2. doi: 10.1176/appi.ajp.162.11.2180

22. Roese, NJ, and Epstude, K. The functional theory of counterfactual thinking: new evidence, new challenges, new insights. In J. M. Olson (Ed.). *Adv Exp Soc Psychol*. (2017) 56:1–79. doi: 10.1016/bs.aesp.2017.02.001

23. Zeelenberg, M, van Dijk, WW, Manstead, ASR, and van der Pligt, J. On bad decisions and disconfirmed expectancies: the psychology of regret and disappointment. *Cognit Emot*. (2010) 14:521–41. doi: 10.1080/026999300402781

24. Jollant, F, Lawrence, NL, Olié, E, Guillaume, S, and Courtet, P. The suicidal mind and brain: a review of neuropsychological and neuroimaging studies. *World J Biol Psychiatry*. (2011) 12:319–39. doi: 10.3109/15622975.2011.556200

25. Camille, N, Coricelli, G, Sallet, J, Pradat-Diehl, P, Duhamel, JR, and Sirigu, A. The involvement of the orbitofrontal cortex in the experience of regret. *Science*. (2004) 304:1167–70. doi: 10.1126/science.1094550

26. Camille, N, Pironti, VA, Dodds, CM, Aitken, MR, Robbins, TW, and Clark, L. Striatal sensitivity to personal responsibility in a regret-based decision-making task. *Cogn Affect Behav Neurosci*. (2010) 10:460–9. doi: 10.3758/CABN.10.4.460

27. Ai, H., van Tol, M. J., Marsman, J. C., Veltman, D. J., Ruhe, H. G., and van der Wee, N. J. A., . . . Aleman, A. (2018). Differential relations of suicidality in depression to brain activation during emotional and executive processing. *J Psychiatr Res*, 105, 78–85. doi:10.1016/j.jpsychires.2018.08.018

28. Zeelenberg, M, van Dijk, WW, and Manstead, AS. Regret and responsibility resolved? Evaluating Ordonez and Connolly's (2000) conclusions. *Organ Behav Hum Decis Process*. (2000) 81:143–54. doi: 10.1006/obhd.1999.2865

29. Dombrovski, AY, Clark, L, Siegle, GJ, Butters, MA, Ichikawa, N, Sahakian, BJ, et al. Reward/punishment reversal learning in older suicide attempters. *Am J Psychiatr*. (2010) 167:699–707. doi: 10.1176/appi.ajp.2009.09030407

30. Beck, AT, Kovacs, M, and Weissman, A. Assessment of suicidal intention: the scale for suicide ideation. *J Consult Clin Psychol*. (1979) 47:343–52. doi: 10.1037/0022-006X.47.2.343

31. Beck, AT, Steer, RA, Ball, R, and Ranieri, W. Comparison of Beck depression inventories -IA and -II in psychiatric outpatients. *J Pers Assess*. (1996) 67:588–97. doi: 10.1207/s15327752jpa6703_13

32. Shek, DT. The Chinese version of the state-trait anxiety inventory: its relationship to different measures of psychological well-being. *J Clin Psychol*. (1993) 49:349–58. doi: 10.1002/1097-4679(199305)49:3<349::aid-jclp2270490308>3.0.co;2-j

33. Bernstein, D. P., Stein, J. A., Newcomb, M. D., Walker, E., Pogge, D., and Ahluvalia, T., . . . Zule, W. Development and validation of a brief screening version of the childhood trauma questionnaire. *Child Abuse Negl*, (2003). 27, 169–190. doi:10.1016/s0145-2134(02)00541-0

34. Treynor, W, Gonzalez, R, and Nolen-Hoeksema, S. Rumination reconsidered: a psychometric analysis. *Cogn Ther Res*. (2003) 27:247–59. doi: 10.1023/A:1023910315561

35. Beck, AT, and Beamesderfer, A. Assessment of depression: the depression inventory. *Mod Probl Pharmacopsychiatry*. (1974) 7:151–69. doi: 10.1159/000395074

36. Patton, JH. Factor structure of the Barratt impulsiveness scale. *J Clin Psychol*. (1995) 51:768–74. doi: 10.1002/1097-4679(199511)51:6<768::AID-JCLP2270510607>3.0.CO;2-1

37. Baskin-Sommers, A, Stuppy-Sullivan, AM, and Buckholtz, JW. Psychopathic individuals exhibit but do not avoid regret during counterfactual decision making. *Proc Natl Acad Sci U S A*. (2016) 113:14438–43. doi: 10.1073/pnas.1609985113

38. Gillan, C. M., Morein-Zamir, S., Kaser, M., Fineberg, N. A., Sule, A., and Sahakian, B. J., . . . Robbins, T. W. Counterfactual processing of economic action-outcome alternatives in obsessive-compulsive disorder: further evidence of impaired goal-directed behavior. *Biol Psychiatry*, (2014). 75: 639–646. doi:10.1016/j.biopsych.2013.01.018

39. Sakamoto, Y, Ishiguro, M, and Kitagawa, G. *Akaike information criterion statistics* D. Reidel Publishing Company (1986) Springer Dordrecht.

40. Shneidman, ES. Suicide, lethality, and the psychological autopsy. *Int Psychiatry Clin*. (1969) 6:225–50. PMID: 5810563

41. Auerbach, RP, Pagliaccio, D, Allison, GO, Alqueza, KL, and Alonso, MF. Neural correlates associated with suicide and non-suicidal self-injury in youth. *Biol Psychiatry*. (2021) 89:119–33. doi: 10.1016/j.biopsych.2020.06.002

42. Hirsch, JK, Duberstein, PR, Conner, KR, Heisel, MJ, Beckman, A, Franus, N, et al. Future orientation and suicide ideation and attempts in depressed adults ages 50 and over. *Am J Geriatr Psychiatry*. (2006) 14:752–7. doi: 10.1097/01.JGP.0000209219.06017.62

43. Wheeler, EZ, and Fellows, LK. The human ventromedial frontal lobe is critical for learning from negative feedback. *Brain*. (2008) 131:1323–31. doi: 10.1093/brain/awn041

44. Bault, N., di Pellegrino, G., Puppi, M., Opolczynski, G., Monti, A., and Braghittoni, D., . . . Coricelli, G. Dissociation between private and social counterfactual value signals following ventromedial prefrontal cortex damage. *J Cogn Neurosci*, (2019). 31:639–656. doi:10.1162/jocn_a_01372

45. Brown, VM, Wilson, J, Hallquist, MN, Szanto, K, and Dombrovski, AY. Ventromedial prefrontal value signals and functional connectivity during decision-making in suicidal behavior and impulsivity. *Neuropsychopharmacology*. (2020) 45:1034–41. doi: 10.1038/s41386-020-0632-0

46. Wenzel, A, Berchick, ER, Tenhave, T, Halberstadt, S, Brown, GK, and Beck, AT. Predictors of suicide relative to other deaths in patients with suicide attempts and suicide ideation: a 30-year prospective study. *J Affect Disord*. (2011) 132:375–82. doi: 10.1016/j.jad.2011.03.006

47. Dombrovski, AY, Hallquist, MN, Brown, VM, Wilson, J, and Szanto, K. Value-based choice, contingency learning, and suicidal behavior in mid- and late-life depression. *Biol Psychiatry*. (2019) 85:506–16. doi: 10.1016/j.biopsych.2018.10.006

48. Richard-Devantoy, S, Olie, E, Guillaume, S, Bechara, A, Courtet, P, and Jollant, F. Distinct alterations in value-based decision-making and cognitive control in suicide attempters: toward a dual neurocognitive model. *J Affect Disord*. (2013) 151:1120–4. doi: 10.1016/j.jad.2013.06.052

49. Loas, G, Lefebvre, G, Rotsaert, M, and Englert, Y. Relationships between anhedonia, suicidal ideation and suicide attempts in a large sample of physicians. *PLoS One*. (2018) 13:e0193619. doi: 10.1371/journal.pone.0193619

50. Winer, ES, Nadorff, MR, Ellis, TE, Allen, JG, Herrera, S, and Salem, T. Anhedonia predicts suicidal ideation in a large psychiatric inpatient sample. *Psychiatry Res*. (2014) 218:124–8. doi: 10.1016/j.psychres.2014.04.016

51. Yang, X, Daches, S, George, CJ, Kiss, E, Kapornai, K, Baji, I, et al. Autonomic correlates of lifetime suicidal thoughts and behaviors among adolescents with a history of depression. *Psychophysiology*. (2019) 56:e13378. doi: 10.1111/psyp.13378

52. Yang, X, Liu, S, Wang, D, Liu, G, and Harrison, P. Differential effects of state and trait social anhedonia on suicidal ideation at 3-months follow up. *J Affect Disord*. (2020) 262:23–30. doi: 10.1016/j.jad.2019.10.056

53. Beck, AT, Brown, G, Berchick, RJ, Stewart, BL, and Steer, RA. Relationship between hopelessness and ultimate suicide: a replication with psychiatric outpatients. *Am J Psychiatr*. (1990) 147:190–5. doi: 10.1176/ajp.147.2.190

54. Fitzgerald, ML, Kassir, SA, Underwood, MD, Bakalian, MJ, Mann, JJ, and Arango, V. Dysregulation of striatal dopamine receptor binding in suicide. *Neuropsychopharmacology*. (2017) 42:974–82. doi: 10.1038/npp.2016.124

55. Pettorruso, M., d'Andrea, G., Martinotti, G., Cocciolillo, F., Miuli, A., and Di Muzio, I., . . . Camardese, G. Hopelessness, dissociative symptoms, and suicide risk in major depressive disorder: clinical and biological correlates. *Brain Sci*, (2020). 10:519. doi:10.3390/brainsci10080519

56. Elledge, D, Zullo, L, Kennard, B, Diederich, A, Emslie, G, and Stewart, S. Refinement of the role of hopelessness in the interpersonal theory of suicide: an exploration in an inpatient adolescent sample. *Arch Suicide Res*. (2021) 25:141–55. doi: 10.1080/13811118.2019.1661896

frontiers | Frontiers in Psychiatry

TYPE Original Research
PUBLISHED 14 March 2023
DOI 10.3389/fpsyt.2023.1137704

Check for updates

OPEN ACCESS

EDITED BY
Hans Albert Braun,
University of Marburg,
Germany

REVIEWED BY
Amanda E. Hernan,
Nemours Children's Health,
United States
Symon M. Kariuki,
KEMRI Wellcome Trust Research Programme,
Kenya

*CORRESPONDENCE
Guoming Luan
✉ luangm3@163.com
Qingyun Wang
✉ nmqingyun@163.com

SPECIALTY SECTION
This article was submitted to Computational Psychiatry, a section of the journal Frontiers in Psychiatry

RECEIVED 04 January 2023
ACCEPTED 21 February 2023
PUBLISHED 14 March 2023

CITATION
Fan D, Wu H, Luan G and Wang Q (2023) The distribution and heterogeneity of excitability in focal epileptic network potentially contribute to the seizure propagation.
Front. Psychiatry 14:1137704.
doi: 10.3389/fpsyt.2023.1137704

The distribution and heterogeneity of excitability in focal epileptic network potentially contribute to the seizure propagation

Denggui Fan[1], Hongyu Wu[1], Guoming Luan[2]* and Qingyun Wang[3]*

[1]School of Mathematics and Physics, University of Science and Technology Beijing, Beijing, China, [2]Epilepsy Center, Sanbo Brain Hospital, Capital Medical University, Beijing, China, [3]Department of Dynamics and Control, Beihang University, Beijing, China

Introduction: Existing dynamical models can explain the transmigration mechanisms involved in seizures but are limited to a single modality. Combining models with networks can reproduce scaled epileptic dynamics. And the structure and coupling interactions of the network, as well as the heterogeneity of both the node and network activities, may influence the final state of the network model.

Methods: We built a fully connected network with focal nodes prominently interacting and established a timescale separated epileptic network model. The factors affecting epileptic network seizure were explored by varying the connectivity patterns of focal network nodes and modulating the distribution of network excitability.

Results: The whole brain network topology as the brain activity foundation affects the consistent delayed clustering seizure propagation. In addition, the network size and distribution heterogeneity of the focal excitatory nodes can influence seizure frequency. With the increasing of the network size and averaged excitability level of focal network, the seizure period decreases. In contrast, the larger heterogeneity of excitability for focal network nodes can lower the functional activity level (average degree) of focal network. There are also subtle effects of focal network topologies (connection patterns of excitatory nodes) that cannot be ignored along with non-focal nodes.

Discussion: Unraveling the role of excitatory factors in seizure onset and propagation can be used to understand the dynamic mechanisms and neuromodulation of epilepsy, with profound implications for the treatment of epilepsy and even for the understanding of the brain.

KEYWORDS

focal epilepsy, network excitability, heterogeneity, synchronous seizure propagation, small-world and free-scale network

1. Introduction

Epilepsy is one of the most common neurological disorders, affecting approximately 65–70 million people worldwide (1). Seizures are usually caused by an imbalance of excitatory and inhibitory cortical neuronal cells (2, 3), and are clinically manifested by massive synchronized periodic discharges based on EEG (Electroencephalogram) (4, 5). Neuronal excitability is associated with a variety of factors, such as microbiota (6), proteases, and glial cells (7). These physiological factors influence neuronal gene expression, morphological development, and the cellular activity (8). As factors that directly or indirectly affect the neuronal excitability expression, they are also associated with epileptic seizures (9). Based on the phenomenon of abnormal excitatory-inhibitory imbalances expressed in epilepsy, non-invasive epilepsy treatments often target on modulating the excitability. For example, antiepileptic drugs can reduce the excitability of neurons by acting on ion channels or indirectly acting on ion channels through neurotransmitter receptors (10), and Deep Brain Stimulation(DBS) can produce excitatory or inhibitory fields thus regulating the state imbalance of nerve cells. However, due to the complicated causes of epilepsy, some patients are resistant to the drugs (11), and still requires sufficient theory and practice to refine the stimulation targets and stimulation patterns. In addition, invasive surgical treatment not only requires a delicate preoperative evaluation but also carries the risk of postoperative paralysis, aphasia and even treatment ineffectiveness. Therefore, understanding the triggers of seizures and the mechanisms of neurophysiological rules governing the development of epileptic brain dynamics may provide theoretical support for epilepsy treatment and may even help us to further understand the brain.

The brain is a highly dynamic system, and human thoughts and memories as well as mechanical movements are controlled and operated by this central system of the brain (12). When any of the activity mechanisms within the brain become abnormal or disrupted, the corresponding brain disorders arise. The pyramidal cells of the neuronal cortex receive either excitatory or inhibitory synaptic potentials and generate extracellular currents (13), they will be detected by tools such as EEG or MEG when many continuously arranged neuronal cells discharge together. Epilepsy is caused by a large number of neuronal cells with hyper-synchronous discharges, and the apparently observable switching of electrical signal patterns during seizures has attracted extensive researchers' attention. The brain is a nonlinear dynamical system, and increasingly mathematical dynamical models have been applied to study and explain the mechanisms of this state transitions (14, 15). For example, models such as Hodgkin-Huxley(HH) (16), Morris-Lecar(ML) (17) elaborate the association of action potential generation with sodium and potassium ions; these describe the behavior of individual neurons at the microscopic level. Models such as Neuron Mass Model (NMM). (18–20) are also included to describe the overall properties of a population of neurons at the macroscopic level, which can better reflect the physiological significance. Typically, a change in the stability of a model caused by a low-dimensional attractor bifurcation in some of the autonomous parameters in the model can induce a seizure-like state of activity. The typical high-frequency rapid discharges (70-120 Hz) that can be recorded at the onset of a seizure and equally accompanied by some low-frequency discharges (β rhythm and γ rhythm, 20–40 Hz) (21). In some cases, some of the parameters in the model can act as control roles for excitability controlling and can provide a rough depiction of the neural field information in a particular state of the brain, simulating abnormal brain firing. Mature model representations and studies have presented us with some of the mechanisms of brain activity, and therefore such models containing excitability information can be used to study the phenomenon of known epileptic hyperexcitability discharges. Besides, there is a separation of time scales during seizures (22), its recurrent nature also suggesting the existence of a larger time scale of epilepsy such as months, years, etc. This indicates that we cannot ignore the differences and associations between different time-scale variables during our modeling process.

Computational models of epilepsy have rapidly advanced and various dynamic mechanisms within the brain can be revealed through computational models. Due to the diverse pathogenesis of epilepsy, different physiological regions result in similar clinical seizure symptoms (23). From these complex physiological mechanisms, common pathways of epilepsy expression can be identified, and such common pathways involve large brain networks. Simplified dynamical models represent only a single modality, and from a dynamic perspective, structural networks characterizing the connectivity of neuronal circuits are often needed to reflect firing activity close to the real physiological mechanisms. Therefore, a combination of dynamical models and brain networks is required to represent the dynamic evolutionary processes more effectively at the whole brain level. It has been established that different network structures embedded in the model lead to different network states (24, 25), and the overall network structure inevitably affects the pattern of information flow traveling through the network. Brain network as a heterogeneous network, with this pattern also related to the properties of each node, which is supported by the interaction of network structure and node excitability distribution (26). The whole brain structural network seems to be considered in most studies where network factors are analyzed, and subnetworks or local networks are mostly considered for their functionality. It is not clear what role the connectivity patterns or nodal properties within their underlying epileptic networks play in triggering the widespread spread of seizures in focal epilepsy. Therefore, a qualitative analysis of our dynamical models in specific structures is necessary.

In this article, we use the model proposed by Jirsa (27), which is a timescale separated model that can separately simulate different types of epileptic-like seizure signals. We simulated a fully connected network model consisting of 100 nodes, in which highly excitatory nodes are considered as "lesion points," which convey excitatory information in the brain. Notably, the focal subnetworks of these focal points are connected to each other in a specific connection pattern with prominent connection strength and without disrupting the fully connected form of the original network. We analyzed the effects of these prominently connected focal nodes on the network model under different structures, different degrees of excitability, and different degrees of excitatory heterogeneity, hoping to provide theoretical support for the mechanism of focal epilepsy generation and focal to bilateral seizures.

2. Models and methods

2.1. Epileptor model

In this paper, we computationally explore the influence factors of seizure propagation of the focal epilepsy network based on the epilepsy oscillator model proposed by Jirsa (27). The model is given as follows:

$$\dot{x}_1 = y_1 - f_1(x_1, x_2) - z + I_1$$

$$\dot{y}_1 = 1 - 5(x_1)^2 - y_1$$

$$\dot{z} = \frac{1}{\tau_o}\left(4(x_1 - x_0) - z\right) \tag{1}$$

$$\dot{x}_2 = -y_2 + x_2 - (x_2)^3 + I_2 + 0.002g(x_1) - 0.3(z - 3.5)$$

$$\dot{y}_2 = \frac{1}{\tau_2}\left(-y_2 + f(x_1, x_2)\right)$$

where

$$f_1(x_1, x_2) = \begin{cases} x_1^3 - 3x_1^2 & x_1 < 0 \\ \left(x_2 - 0.6(z-4)^2 x_1\right) & x_1 \geq 0 \end{cases}$$

$$f_2(x_1, x_2) = \begin{cases} 0 & x_1 < -0.25 \\ 6(x_2 + 0.25) & x_1 \geq -0.25 \end{cases} \tag{2}$$

and

$$g(x_1) = \int_{t_0}^{t} e^{-^3(t-\tau)} x_1(\tau)dt \tag{3}$$

This model includes three groups of variables with different time scales. x_1, x_2 are responsible for generating fast oscillations, related to the potential activity of the neuronal membrane, with the shortest time scale. y_1, y_2 are responsible for generating SWE (sharp-wave events) and interictal spikes, with a slower time scale τ_2 compared to the first group of variables, simulating the membrane potential. The variable z has the largest time scale and represents the slowly varying permittivity variable responsible for guiding the entire system. During epileptic-like seizures, z is associated with slowly changing processes outside the cell, such as ion levels, energy metabolism and oxygen content, etc. In this model, $x_1 + x_2$ can be used to represent the electrographic signatures of a SLE (Seizure like events).

In addition to the interaction between the fast and slow nervous system in the model, it is also coupled through the permittivity variable z, as shown in Figure 1A (28). The dielectric coefficients are considered to be correlated with excitability and control the onset state of the model. Figure 1B (28) gives a bifurcation diagram of the fast variable x_1 with regards to the variable z. When the dielectric coefficient goes from large to small through the SNIC (Saddle-node on invariant circle), the model transitions from the interictal to the ictal state. Conversely, when the dielectric coefficient goes from small to large through the HB (Homoclinic bifurcation), the model transitions from the ictal state to the interictal state. This bistable mechanism leads to the existence of an "epileptic element" x_0 in the model that controls state switching, which can be used as the threshold to control the onset of the model (according to current model, $x_{threshold} = -2.05$). When $x_0 < x_{threshold}$ the system stays at a stable fixed point and does not generate seizures, while when the $x_0 > x_{threshold}$, the system will transit to the seizure phase. What is more, the excitability of one node depends on the distance between x_0 and $x_{threshold}$. The healthy node may also be recruited to present a seizure state under external perturbations if the distance is too close (Figure 1C).

2.2. Whole brain network model

To investigate the seizure effects of focal epilepsy on a whole-brain scale, we modeled the brain as a network. Individual brain regions or clusters of neuronal cells can be taken as nodes, and the connections between them are mapped to become the edges of the network. In the dynamical model, each epileptic oscillator can be seen as part of a brain region, and nodal connections can be implemented by coupling in the model.

It has been indicated that fast coupling through synaptic or gap connections does not induce qualitative variations in slow time-scale behavior (29), thus the multi-timescale model of epilepsy with recurrent seizures needs to take into account the slow dielectric coefficients containing cellular parameters (30). This oscillator model is a phenomenological model presenting epilepsy-like activity and has less direct connection to the biophysiological mechanisms embedded in the real human brain. Starting from the phenomenology, the permittivity variable z on the slow time scale is coupled with linear inhibition of the fast and slow subsystems and negative feedback coupling to SLE. In the case of multiple nodes discharging simultaneously, the discharge of node j can be conveyed to the vicinity of node i through axonal transmission, which perturbs the dynamical state of node i . The axonal connections that play the role of axonal transmission are represented in the form of structural connections. In a whole-brain network, all nodes can be coupled using a permittivity variable z that represents a process external to the cell. Thus, a model of a whole-brain network formed by multiple epileptic oscillators is as follows:

$$\dot{x}_{1,i} = y_{1,i} - f_1(x_{1,i}, x_{2,i}) - z_i + I_1$$

$$\dot{y}_{1,i} = 1 - 5(x_{1,i})^2 - y_{1,i}$$

$$\dot{z}_i = \frac{1}{\tau_o}\left(4(x_{1,i} - x_0) - z_i - \sum_{j=1}^{n} S_{ij}(x_{1,i} - x_{1,j})\right) \tag{4}$$

$$\dot{x}_{2,i} = -y_{2.i} + x_{2,i} - (x_{2,i})^3 + I_2 + 0.002g(x_{2,i}) - 0.3(z_i - 3.5)$$

$$\dot{y}_{2,i} = \frac{1}{\tau_2}\left(-y_{2.i} + f(x_{1,i}, x_{2,i})\right)$$

where S_{ij} represents the degree of connectivity between individual nodes, which can usually be represented by the structural

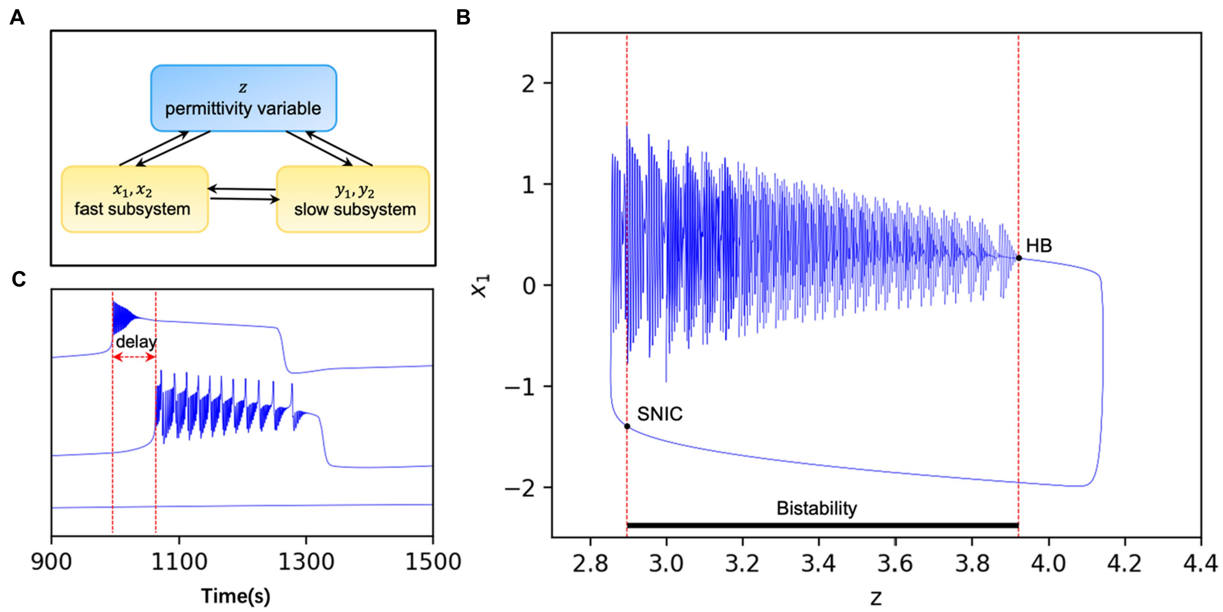


FIGURE 1
(A) Pattern diagram of the model. x_1, x_2 are the fast subsystems and y_1, y_2 are the slow subsystems. In addition to their direct interaction, the two subsystems can also interact with each other indirectly through the extracellular environmental, i.e., the permittivity variable z. **(B)** The system is bistable when the permittivity variable z lies between two bifurcation points and loses stability when the stable limit ring passes through the HB or the stable point passes through SNIC. The whole system varies periodically with time modified from Guo et al. (28). **(C)** Discharges of nodes with different excitability in the same network. When $x_0 > x_{threshold}$, the model is in the seizure state and the oscillator will switch periodically between the seizure and interseizure periods (upper panel). When $x_0 < x_{threshold}$, the oscillator in the non-ictal state switches between the onset and interictal states due to the perturbations such as network connection and noise, and there is often a time delay with the oscillator in the onset state (middle panel). Also, when $x_0 < x_{threshold}$, the model is in the non-seizure state and the oscillator is stable for a long time without perturbations (lower panel).

TABLE 1 Model default parameter.

Parameter	Value	Meaning
I_1	3.1	Current of fast subsystem
I_2	0.45	Current of slow subsystem
τ_0	2,857	Time scale of the permittivity variable
τ_2	10	Time scale of the permittivity variable
γ	0.01	Time constant in function $g(x)$

connectivity matrix of the brain. The model is simulated by fourth order Runge–Kutta, and all parameters in the model are shown in Table 1.

2.3. Network structure and excitatory heterogeneity of seizure nodes

In our work, to explore the seizure propagation of focal epilepsy, we have considered several factors that may influence the outcome of propagation. The first is the connectivity structure of the lesion nodes. Complex network theory provides a rich perspective and tool for brain network studies (31–33), and classical network models such as random networks often have their unique properties that can be used to depict rich brain network connections. Several classical complex network models including random networks, small-world networks, and scale-free networks are introduced into the network dynamics model in this paper. We built a special fully connected network. First, the strength of connections in this network is inversely proportional to the paths between nodes pairs, then the connections between groups of excitatory nodes (which can be considered as lesion nodes) were strengthened to form a specific network model structure individually. In a whole perspective, the network remains a fully connected network with the lesion nodes are prominently connected. This situation can be seen as a special network structure embedded in the original fully connected network, as shown in Figure 2, in which connection strength is of significant differences. In this way, we obtain a connectivity matrix S_{ij} of the fully connected network. Besides, the proportion of focal nodes is also taken into account. Different lesion proportion implies different scales of lesion networks, which is one of the influencing factors that we cannot ignore.

The topological connectivity and the scale of the network can be considered as the “physical properties” of a network, where each node is simulated by a dynamic model, and the variable z in each model represents the degree of excitability of the node, controlled by x_0. The difference in excitability of each node in the network can be considered as the unique “intrinsic property” of each network. We replace a set of excitatory nodes with x_0 following normal distribution into the focal network:

$$x_{0,i} \sim N\left(\mu, \sigma^2\right), x_{0,i} > x_{threshold}, \\ for\ i = 1, 2, \ldots\ldots, n, i = lesion\ node \tag{5}$$

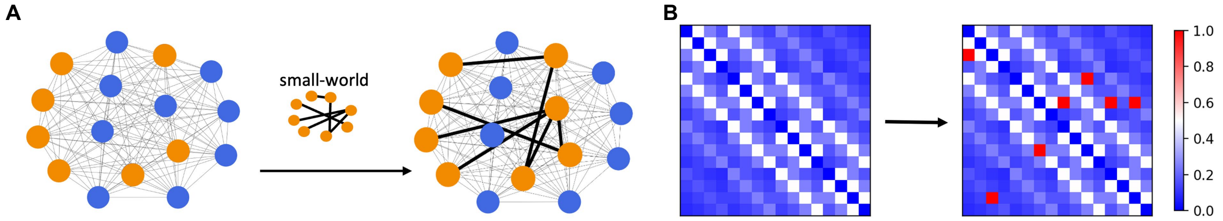


FIGURE 2
Diagram of network connection structure. **(A)** The strength of connections between nodes in a fully connected network is inversely proportional to the between-node path. The connection pattern between lesion nodes (orange) is topologically specific (small-world connection as an example). The strength of the connection between the lesion nodes was significantly greater than the remaining connections **(B)** The connectivity strength between lesion nodes is extremely prominent, much higher than that between non-lesion nodes, and this matrix can be used as the connectivity matrix *Sij* in the model.

The excitability and heterogeneity of nodes can be expressed as σ and σ , respectively. The excitatory nodes should not be too close to the threshold and the heterogeneity should not be too large, otherwise some nodes may be included in the non-epileptogenic zone.

3. Results

3.1. Whole brain network connectivity mechanisms underlying the consistent discharges

In the epileptor model, different permittivity coefficients, i.e., variable z, guides the system into different states. And the epileptogenic factor x_0 included in z can be used as the main parameter to control the degree of excitability of the node. x_0 located on the left and right sides of the $x_{threshold}$ causes the system to be in a non-oscillatory state and an oscillatory state, respectively, where the oscillatory state can be considered as the seizure state. In a system with individual node, the model is governed by a single excitability index x_0 . In the network model, the interactions between nodes implies a diversity of node states. This multi-state is not only determined by the initial diverse excitability of the nodes, but the connectivity between nodes embedded in z also influences the state of the nodes in some way. We set some of the nodes in the multi-node network as excitatory nodes and the rest as non-excitatory nodes. We found that when the node network is sparsely connected, due to the presence of excitatory nodes, part of non-excitatory nodes also exhibits state switching, but the overall excitatory synchronization of the network is weak (Figure 3A), but when the node network is fully connected, all the non-excitable nodes are also converted to a “delayed onset” oscillatory state in the network model due to the overlapping of node interactions,and most of the nodes have high excitatory synchronization (Figure 3B). However, without the existence of excitatory nodes, full connectivity between nodes cannot directly cause state switching in some nodes either (Figure 3C). We speculate that the primary condition controlling the dynamical behavior of brain regions or neuronal cells within the brain is their own physiological situation, but the information transfer and interaction relationship between individual units is also a part that cannot be ignored.

3.2. The effects of focal network size and excitability patterns on the epileptic seizure periods

Epileptic seizures formally exhibit large-scale periodic coherent discharges. In the results of model simulations, we can also observe periodic changes in the fast and slow subsystems and variable z of excitatory nodes. The multiple time scales involved in epilepsy have been of wide interest, different time scales involve different physiological dynamic behaviors. For epilepsy which may persist with recurrent seizures over a long period of time, time plays an important role, with short periods implying frequent and continuous seizures, which pose a great challenge to the patient himself and to the treatment. Long periods may offer the possibility of interrupting the process of the disease. The electrophysiological mechanisms underlying the switch between ictal and interictal periods in such periodic discharges may conceal the triggering of seizures. In this work, to explore the factors that influence the period of epileptic discharges in a known dynamic background, we considered the network situation in a multi-node model and the excitability of the network nodes. We found that both the proportion of excitatory nodes and their epileptogenic factor x_0 influence the period of the synchronous oscillation of the nodes. Holding the remaining factors constant, the period of oscillation of the network is negatively correlated with both the averaged x_0 (i.e., average excitability μ) and the proportion of excitatory nodes. When x_0 is located in the excitatory region, the greater the distance from the threshold [Figures 4B,C (lower)], the greater the proportion of excitatory nodes [Figures 4A,C (middle)], the shorter the period, the more frequent the state switching of the nodes. And it is not affected by the heterogeneity of node excitability [Figure 4C (upper)]. This may imply that the excitability level of nodes plays an essential role in the network model, and either the change in excitability of a single node itself or the accumulation of multiple similarly excitable nodes will change the overall excitability of the network model, which will be reflected in the periodicity and frequency of seizures.

3.3. The secondary effect of focal network topology on its functional activity

In the field of brain network research, statistical relationships between signals are often used to build functional networks to

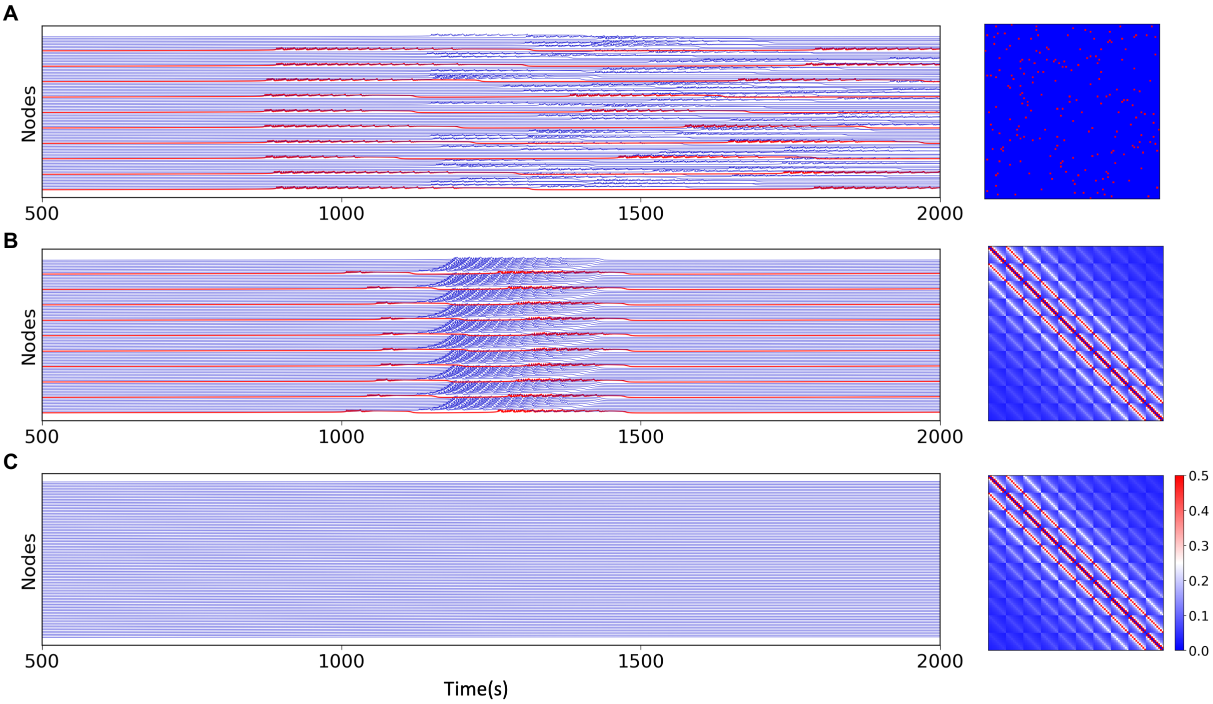


FIGURE 3
Simulation results for the network model with different network connections and different excitations, the connection matrix is shown on the right. **(A)** Simulation results of the multi-node network model under the random network, where the excitatory nodes (red) can have state switching and a portion of the non-excitatory nodes remain in the stable state. **(B)** Simulation results of the multi-node model with fully connected network. Except for excitatory nodes (red), all non-excitatory nodes generate state switching, which is slightly delayed than excitatory nodes. **(C)** Simulation results of the model without excitatory nodes under fully connected network connections, all nodes do not have state switching.

investigate the functional coherence of individual brain regions or nodes. As mentioned previously we constructed a fully connected network with lesion nodes specifically connected and strongly connected. After obtaining multiple sets of simulated signals for the same lesion node proportion of the network model, we calculated the Pearson correlation between the signals and used them as the edges of the network to construct a functional network of simulated signals. In this way, we observed the characteristics of the lesion nodes structurally and functionally. We preserved the top 27.5% of the functional connectivity strength to visualize the structure of the functional network. As shown in Figure 5, the average degree of lesion nodes in the functional network correlates with the heterogeneity in the excitability of the lesion nodes, regardless of the connectivity pattern. The functional network after sparing is preserved as strongly connected, with each connection representing a high correlation between signals. When σ = 0, the lesion nodes are homogeneous, and the average degree of all lesion points is maintained around the lesion proportion. Larger σ represents a greater degree of heterogeneity in node excitability, while the potential average activity level (average degree) of the corresponding lesion cluster is negatively correlated with σ, and the connectivity of the lesion cluster becomes smaller as σ increases. However, structural changes in different connectivity patterns under the same type of network did not have a dramatic effect on this trend overall (Figure 5). This implies that excitability in the network remains the dominant factor influencing the state of the system. We noted subtle effects from changes in network structure, but they remained a secondary condition compared to excitability.

4. Discussion

Seizures involve abnormalities related to ion channels and synaptic function, and the brain excitation/inhibition circuits develop a dysfunction, which in turn leads to an imbalance of excitation and inhibition in the brain system, usually manifesting as hyperexcitability (34, 35). Some of the disorders caused by excitability-related elemental abnormalities are also accompanied by the generation of epilepsy (36, 37). The process of using DBS for drug-resistant epilepsy is to alter the activity of local field potentials and the excitability of brain networks by remote thalamic stimulation or direct cortical stimulation (38). It is thus clear that excitability is a never-ending subject in the field of epilepsy. However, epilepsy remains a challenge in modern medicine, with its complex temporal and spatial scales, and the seizure mechanisms involved have not been fully revealed. Existing ideas include recording a series of imaging data before and after a clinical seizure, which allows to analyze and predict the seizure and propagation of epilepsy, etc. (39–41). Data analysis is difficult to avoid the specificity brought by individual data, and models can fill the missing part of data analysis. There is a rich electrophysiological mechanism behind the operations of the brain, and some of these transitions can be well reproduced by existing dynamical models, and mature nonlinear

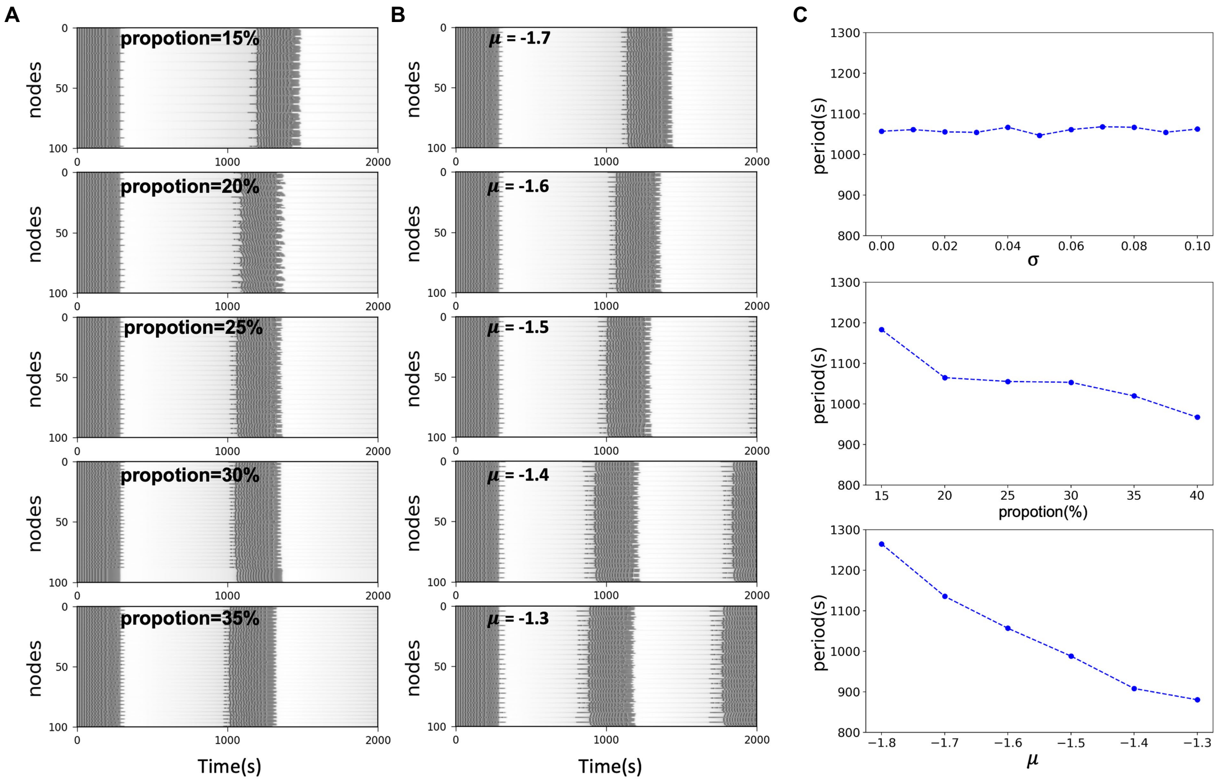


FIGURE 4
(A) Simulated time series of network model with different excitatory node proportions. In the same scale network, the larger the excitability proportion, the shorter the system oscillation period. **(B)** Simulated time series of network model with the same proportion and different excitability level μ. In the same scale network, the closer the distance between $x0$ and the threshold, the shorter the oscillation period of the system. **(C)** System oscillation period curve with respect to the excitability heterogeneity of network nodes, network size and the network excitability level, respectively.

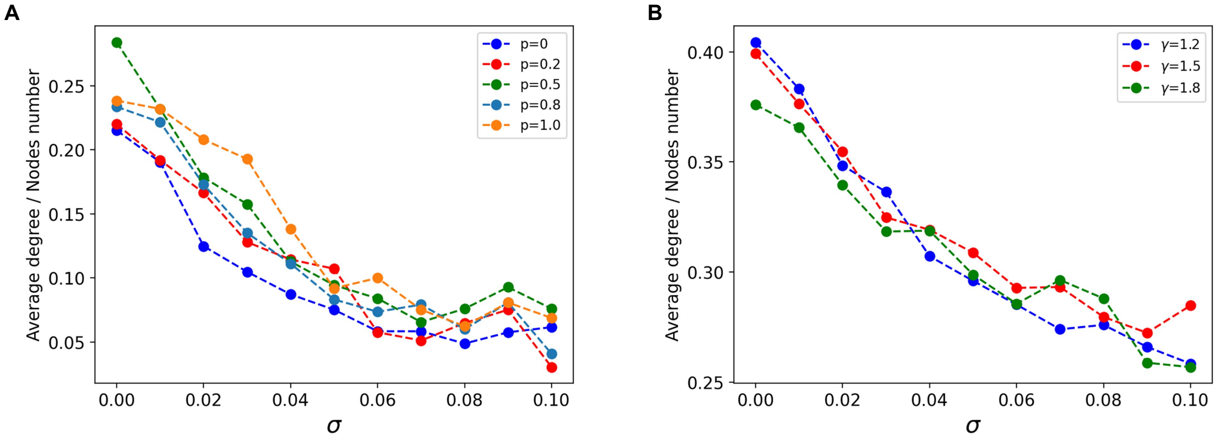


FIGURE 5
Curves of the relationship between the average degree of lesion node groups and excitatory heterogeneity in the functional network. **(A)** In the strongly connected network of lesion nodes, the higher the node heterogeneity, the smaller the average degree of nodes, and the regular connection (p = 0), small world connection (0 < p < 1) and random connection (p = 1) show the same law. **(B)** In the scale-free strong connection network of focal nodes (γ is the power-law exponent of scale-free network), the higher the node heterogeneity, the smaller the average degree of nodes, and the scale-free network structure pattern has little influence.

dynamical theories can be combined with models to explain some brain-like phenomena. A strong and distinct state switching mechanism exists in epilepsy and is accompanied by a certain periodicity. Combining excitatory factors with computational models to explore the mechanisms of seizures can provide theoretical support for our treatment and study of epilepsy.

In this article, we first analyze the nonlinear dynamical features in the model to explore the mechanisms of abnormal discharges.

We found that this abnormal discharge is controlled by a set of bifurcations, with seizures starting at the saddle-point bifurcation and ending at the homozygous bifurcation. The simple model generates abnormal discharges on the premise that to control the model lies within the excitatory region. We also found that excitability is the main factor affecting the model state in a mutually coupled network model. Non-excited nodes have a certain probability to produce a delayed discharge behavior over excited nodes in the case of node coupling overlapping. Such a delayed spontaneous discharge phenomenon helps us to understand the direction of information flow in epileptic networks. In addition to this, we control the distance of the parameters of node excitability and the proportion of excitable nodes in the network, which significantly affect the period of system discharge. This implies that changes in excitability in either degree or extensity affect the system state. The effects of altered system excitability are reflected in both microbiological and macroscopic computational models. Complex systems often contain multi-element interactions, and multi-element excitatory heterogeneity has been similarly shown to play with a role in the propagation of epilepsy when the overall excitability of the network system is constant (42). In our work, the lower the excitability heterogeneity, the stronger the association between clusters of excitatory nodes, which is reflected in the functional network of lesion nodes after model simulation. We suggest that excitability is in a primary position compared to other factors including network coupling and network structure, and that excitability can produce effects on the system in multiple dimensions.

In our whole-brain network, network coupling is not based on structural connectome, but rather a structural network with prominent lesion connections, which theoretically establishes a deeper understanding of the relationship between excitability and the system. The advantage of this is a clearer understanding of the coupling information, and the disadvantage is the lack of physiological information about the real situation of the brain. Currently, in the context of such fully connected networks, there is no good correspondence between simulated signals and known specific network structures for analysis, which is our later effort. It is expected that the dynamic flow of neural information in specific network structures can be revealed. During the model simulation, excitability shows its main role in controlling the state of the system. In the case of epilepsy, such results are inevitable. Therefore, there is a strong need to focus the perspective on potential influences beyond excitability to provide more diverse theoretical support for seizure mechanisms as well as modeling. In conclusion, this study analyzed how excitability parameters in the model affect the dynamic switching as well as the intrinsic properties of the network system in different perspectives by modeling the dynamics and parameter modulation of the epileptic network, reflecting the importance of excitability factors in the epileptic system.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding authors.

Author contributions

DF proposed and supervised the project and contributed to writing the manuscript. HW analyzed the data, performed the experiments, and wrote the manuscript. QW and GL supervised and revised the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This research was supported by the National Natural Science Foundation of China (Grants 12072021 and 11932003) and the Young Teachers International Exchange Growth Program (QNXM20220049).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Hesdorffer DC, Logroscino G, Benn EKT, Katri N, Cascino G, Hauser WA. Estimating risk for developing epilepsy: a population-based study in Rochester, Minnesota. *Neurology*. (2011) 76:23–7. doi: 10.1212/WNL.0b013e318204a36a

2. Moore AK, Weible AP, Balmer TS, Trussell LO, Wehr M. Rapid rebalancing of excitation and inhibition by cortical circuitry. *Neuron*. (2018) 97:1341–55.e6. doi: 10.1016/j.neuron.2018.01.045

3. Heise C, Taha E, Murru L, Ponzoni L, Cattaneo A, Guarnieri FC, et al. eEF2K/eEF2 pathway controls the excitation/inhibition balance and susceptibility to epileptic seizures. *Cereb Cortex*. (2017) 27:2226–48. doi: 10.1093/cercor/bhw075

4. Stafstrom CE, Carmant L. Seizures and epilepsy: an overview for neuroscientists. *Cold Spring Harb Perspect Med*. (2015) 5:a022426. doi: 10.1101/cshperspect.a022426

5. Stevens JR, Lonsbury BL, Goel SL. Seizure occurrence and interspike interval: telemetered electroencephalogram studies. *Arch Neurol*. (1972) 26:409–9. doi: 10.1001/archneur.1972.00490110043004

6. Darch H, McCafferty CP. Gut microbiome effects on neuronal excitability & activity: implications for epilepsy. *Neurobiol Dis*. (2022) 165:105629. doi: 10.1016/j.nbd.2022.105629

7. Verhoog QP, Holtman L, Aronica E, van Vliet EA. Astrocytes as guardians of neuronal excitability: mechanisms underlying epileptogenesis. *Front Neurol*. (2020) 11:591690. doi: 10.3389/fneur.2020.591690

8. Jaworski T. Control of neuronal excitability by GSK-3beta: epilepsy and beyond. Biochimica et Biophysica Acta (BBA)-molecular. *Cell Res*. (2020) 1867:118745. doi: 10.1016/j.bbamcr.2020.118745

9. Bonnet U, Bingmann D, Speckmann EJ, Wiemann M. Small intraneuronal acidification via short-chain monocarboxylates: first evidence of an inhibitory action on over-excited human neocortical neurons. *Life Sci*. (2018) 204:65–70. doi: 10.1016/j.lfs.2018.05.005

10. Rogawski MA, Löscher W. The neurobiology of antiepileptic drugs. *Nat Rev Neurosci*. (2004) 5:553–4. doi: 10.1038/nrn1430

11. Schmidt D, Löscher W. Drug resistance in epilepsy: putative neurobiologic and clinical mechanisms. *Epilepsia*. (2005) 46:858–7. doi: 10.1111/j.1528-1167.2005.54904.x

12. Zatorre RJ, Fields RD, Johansen-Berg H. Plasticity in gray and white: neuroimaging changes in brain structure during learning. *Nat Neurosci*. (2012) 15:528–6. doi: 10.1038/nn.3045

13. Levine DN. Sherrington's "the integrative action of the nervous system": a centennial appraisal. *J Neurol Sci*. (2007) 253:1–6. doi: 10.1016/j.jns.2006.12.002

14. Marten F, Rodrigues S, Suffczynski P, Richardson MP, Terry JR. Derivation and analysis of an ordinary differential equation mean-field model for studying clinically recorded epilepsy dynamics. *Phys Rev E*. (2009) 79:021911. doi: 10.1103/PhysRevE.79.021911

15. Goodfellow M, Schindler K, Baier G. Intermittent spike–wave dynamics in a heterogeneous, spatially extended neural mass model. *NeuroImage*. (2011) 55:920–2. doi: 10.1016/j.neuroimage.2010.12.074

16. Hodgkin AL, Huxley AF. A quantitative description of membrane current and its application to conduction and excitation in nerve. *J Physiol*. (1952) 117:500–4. doi: 10.1113/jphysiol.1952.sp004764

17. Naze S, Bernard C, Jirsa V. Computational modeling of seizure dynamics using coupled neuronal networks: factors shaping epileptiform activity. *PLoS Comput Biol*. (2015) 11:e1004209. doi: 10.1371/journal.pcbi.1004209

18. Wendling F, Bartolomei F, Bellanger JJ, Chauvel P. Epileptic fast activity can be explained by a model of impaired GABAergic dendritic inhibition. *Eur J Neurosci*. (2002) 15:1499–08. doi: 10.1046/j.1460-9568.2002.01985.x

19. Jansen BH, Rit VG. Electroencephalogram and visual evoked potential generation in a mathematical model of coupled cortical columns. *Biol Cybern*. (1995) 73:357–6. doi: 10.1007/BF00199471

20. Jansen BH, Zouridakis G, Brandt ME. A neurophysiologically-based mathematical model of flash visual evoked potentials. *Biol Cybern*. (1993) 68:275–3. doi: 10.1007/BF00224863

21. Worrell GA, Parish L, Cranstoun SD, Jonas R, Baltuch G, Litt B. High-frequency oscillations and seizure generation in neocortical epilepsy. *Brain*. (2004) 127:1496–06. doi: 10.1093/brain/awh149

22. Baud MO, Ghestem A, Benoliel JJ, Becker C, Bernard C. Endogenous multidien rhythm of epilepsy in rats. *Exp Neurol*. (2019) 315:82–7. doi: 10.1016/j.expneurol.2019.02.006

23. Richardson MP. Large scale brain models of epilepsy: dynamics meets connectomics. *J Neurol Neurosurg Psychiatry*. (2012) 83:1238–48. doi: 10.1136/jnnp-2011-301944

24. Benjamin O, Fitzgerald TH, Ashwin P, Tsaneva-Atanasova K, Chowdhury F, Richardson MP, et al. A phenomenological model of seizure initiation suggests network structure may explain seizure frequency in idiopathic generalised epilepsy. *J Math Neurosci*. (2012) 2:1–30. doi: 10.1186/2190-8567-2-1

25. Petkov G, Goodfellow M, Richardson MP, Terry JR. A critical role for network structure in seizure onset: a computational modeling approach. *Front Neurol*. (2014) 5:261. doi: 10.3389/fneur.2014.00261

26. Lopes MA, Junges L, Woldman W, Goodfellow M, Terry JR. The role of excitability and network structure in the emergence of focal and generalized seizures. *Front Neurol*. (2020) 11:74. doi: 10.3389/fneur.2020.00074

27. Jirsa VK, Stacey WC, Quilichini PP, Ivanov AI, Bernard C. On the nature of seizure dynamics. *Brain*. (2014) 137:2210–30. doi: 10.1093/brain/awu133

28. Guo D, Xia C, Wu S, Zhang T, Zhang Y, Xia Y, et al. Stochastic fluctuations of permittivity coupling regulate seizure dynamics in partial epilepsy. *Sci China Technol Sci*. (2017) 60:995–02. doi: 10.1007/s11431-017-9030-4

29. Haken H. *Advanced synergetics: instability hierarchies of self-organizing systems and devices*, vol. *20*. Berlin: Springer Science & Business Media (2012).

30. Proix T, Bartolomei F, Chauvel P, Bernard C, Jirsa VK. Permittivity coupling across brain regions determines seizure recruitment in partial epilepsy. *J Neurosci*. (2014) 34:15009–21. doi: 10.1523/JNEUROSCI.1570-14.2014

31. Supekar K, Menon V, Rubin D, Musen M, Greicius MD. Network analysis of intrinsic functional brain connectivity in Alzheimer's disease. *PLoS Comput Biol*. (2008) 4:e1000100. doi: 10.1371/journal.pcbi.1000100

32. Hagmann P, Kurant M, Gigandet X, Thiran P, Wedeen VJ, Meuli R, et al. Mapping human whole-brain structural networks with diffusion MRI. *PLoS One*. (2007) 2:e597. doi: 10.1371/journal.pone.0000597

33. Eguiluz VM, Chialvo DR, Cecchi GA, Baliki M, Apkarian AV. Scale-free brain functional networks. *Phys Rev Lett*. (2005) 94:018102. doi: 10.1103/PhysRevLett.94.018102

34. Noebels JL. Targeting epilepsy genes. *Neuron*. (1996) 16:241–4. doi: 10.1016/S0896-6273(00)80042-2

35. Babb TL, Pretorius JK, Kupfer WR, Crandall PH. Glutamate decarboxylase-immunoreactive neurons are preserved in human epileptic hippocampus. *J Neurosci*. (1989) 9:2562–74. doi: 10.1523/JNEUROSCI.09-07-02562.1989

36. Bozzi Y, Provenzano G, Casarosa S. Neurobiological bases of autism–epilepsy comorbidity: a focus on excitation/inhibition imbalance. *Eur J Neurosci*. (2018) 47:534–8. doi: 10.1111/ejn.13595

37. Scharfman HE, MacLusky NJ. The influence of gonadal hormones on neuronal excitability, seizures, and epilepsy in the female. *Epilepsia*. (2006) 47:1423–40. doi: 10.1111/j.1528-1167.2006.00672.x

38. Stypulkowski PH, Stanslaski SR, Jensen RM, Denison TJ, Giftakis E. Brain stimulation for epilepsy–local and remote modulation of network excitability. *Brain Stimul*. (2014) 7:350–8. doi: 10.1016/j.brs.2014.02.002

39. Li Y, Liu Y, Cui WG, Guo YZ, Huang H, Hu ZY. Epileptic seizure detection in EEG signals using a unified temporal-spectral squeeze-and-excitation network. *IEEE Trans Neural Syst Rehabil Eng*. (2020) 28:782–4. doi: 10.1109/TNSRE.2020.2973434

40. Rasheed K, Qayyum A, Qadir J, Sivathamboo S, Kwan P, Kuhlmann L, et al. Machine learning for predicting epileptic seizures using EEG signals: a review. *IEEE Rev Biomed Eng*. (2020) 14:139–5. doi: 10.1109/RBME.2020.3008792

41. Wei X, Zhou L, Zhang Z, Chen Z, Zhou Y. Early prediction of epileptic seizures using a long-term recurrent convolutional network. *J Neurosci Methods*. (2019) 327:108395. doi: 10.1016/j.jneumeth.2019.108395

42. Wang C, Chen S, Huang L, Yu L Prediction and control of focal seizure spread: random walk with restart on heterogeneous brain networks. arXiv preprint arXiv:220406939 (2022).

TYPE Original Research
PUBLISHED 27 March 2023
DOI 10.3389/fpsyt.2023.1083015

Check for updates

OPEN ACCESS

EDITED BY
Jianzhong Su,
University of Texas at Arlington, United States

REVIEWED BY
Yanmei Kang,
Xi'an Jiaotong University, China
Wei Lin,
Fudan University, China

*CORRESPONDENCE
Jianwei Shen
✉ xcjwshen@gmail.com

SPECIALTY SECTION
This article was submitted to Computational Psychiatry, a section of the journal Frontiers in Psychiatry

RECEIVED 28 October 2022
ACCEPTED 14 February 2023
PUBLISHED 27 March 2023

CITATION
Wang J and Shen J (2023) Turing instability mechanism of short-memory formation in multilayer FitzHugh-Nagumo network. *Front. Psychiatry* 14:1083015. doi: 10.3389/fpsyt.2023.1083015

Turing instability mechanism of short-memory formation in multilayer FitzHugh-Nagumo network

Junjie Wang[1] and Jianwei Shen[2]*

[1]School of Mathematics and Statistics, Zhengzhou University, Zhengzhou, China, [2]School of Mathematics and Statistics, North China University of Water Resources and Electric Power, Zhengzhou, China

Introduction: The study of brain function has been favored by scientists, but the mechanism of short-term memory formation has yet to be precise.

Research problem: Since the formation of short-term memories depends on neuronal activity, we try to explain the mechanism from the neuron level in this paper.

Research contents and methods: Due to the modular structures of the brain, we analyze the pattern properties of the FitzHugh-Nagumo model (FHN) on a multilayer network (coupled by a random network). The conditions of short-term memory formation in the multilayer FHN model are obtained. Then the time delay is introduced to more closely match patterns of brain activity. The properties of periodic solutions are obtained by the central manifold theorem.

Conclusion: When the diffusion coeffcient, noise intensity *np*, and network connection probability *p* reach a specific range, the brain forms a relatively vague memory. It is found that network and time delay can induce complex cluster dynamics. And the synchrony increases with the increase of *p*. That is, short-term memory becomes clearer.

KEYWORDS

FHN model, short-term memory, multilayer network, Turing pattern, delay, Hopf bifurcation, noise

1. Introduction

In 1952, Alan Hodgkin and Andrew Huxley developed the famous Hodgkin-Huxley (HH) model based on nerve stimulation potential data of squid. Due to the high dimension and computational complexity of the HH model, Richard FitzHugh and J.Nagumo simplified the HH model and established the FHN model. In the actual nerve conduction process, there is a certain time delay in signal transmission, which caused a lot of research on the FHN model with time delay. Wang et al. studied bifurcation and synchronization (1), bifurcation structure (2), Fold-Hopf bifurcation (3), periodic oscillation (4), and global Hopf bifurcation (5) of coupled FHN model with time delay. Yu et al. (6) found that the noise level can change the signal transmission performance in the FHN network, and the delay can cause multiple stochastic resonances. Gan et al. (7) also found that appropriate delay can induce stochastic resonances in FHN scale-free networks and devoted themselves to extending the range of stochastic resonance on complex neural networks. Zeng et al. (8) found that, unlike noise, the system undergoes a phase transition as the time delay increases. Bashkirtseva and Ryashko analyzed the excitability of the FHN model using the stochastic sensitivity function technique and proposed a new method for analyzing attractors (9). In addition, it is found that there are very complex dynamic phenomena in the FHN model. Rajagopal et al. (10) studied chaos and periodic bifurcation diagrams under different excitation currents and found that the dynamic behavior of the nodes alters dramatically after the introduction of Gaussian noise.

Iqbal et al. (11) studied robust adaptive synchronization of a ring-coupled uncertain chaotic FHN model and proposed a scheme to synchronize the coupled neurons under external electrical stimulation. Feng et al. studied the influence of external electromagnetic radiation on the FHN model. And they found that periodic, quasi-periodic, and chaotic motions would occur in different frequency intervals when the external electromagnetic radiation was in the form of a cosine function (12).

In the same year, the HH model was proposed, Turing discovered that a stable uniform state would become unstable under certain conditions in a reaction-diffusion system, which attracted a lot of attention and was introduced into various fields. Liu et al. found that cross-diffusion could lead to Turing instability of periodic solutions (13, 14). Lin et al. analyzed the conditions of Turing-Hopf bifurcation and the spatiotemporal dynamics near the bifurcation point in diffusion neural networks with time delay (15). Mondal et al. studied the dynamical behaviors near the Turing-Hopf bifurcation points of the neural model. And they found that collective behaviors may be related to the generation of some brain pathologies (16). Qu and Zhang studied the conditions required for various bifurcations in the FHN diffusion system under Neumann boundary conditions and extended them to coupled FHN model (17). With the boom of complex networks in recent years, many scholars have begun to study the Turing pattern under the network (18–22). Ren et al. extended these studies to multilayer networks (23, 24). Moreover, Tian et al. investigated pattern and Hopf bifurcation caused by time delay in the Small-World network, Barabasi-Albert free-scale network, and Watts-Strogatz network (25, 26). These studies take the pattern problem to a new level.

Researchers are keen to study some characteristic behaviors of the brain from the perspective of the network because the brain is a complex network system with hierarchical and modular structures (27). Neurons generate complex cluster dynamic behaviors through synaptic coupling to form brain function. Neurons with similar connection patterns usually have the same functional attributes (28). Experiments have shown that neurons far apart can fire simultaneously when the brain is stimulated and that this phenomenon persists when neurons are in the resting state. One of the brain's basic functions is remembering information, which can be a sensory stimulus or a text (29). The principle of memory formation is very complex and is still being explored. A classic view is that the realization of short-term memory in the brain depends on fixed point attractors (30, 31). Memory storage is maintained by the continuous activity of neurons, which persists even after the memory stimulus has been removed (32, 33). Goldman showed the fundamental mechanisms that generate sustained neuronal activity in feedforward and recurrent networks (33). Neurons release neurotransmitters that direct human activity when the brain receives the information. However, due to the noise and the existence of inhibitory neurons, information processing cannot always be synchronized in time, which leads to a certain delay in the recovery time of action potentials (34). And Yu et al. found that the delay will affect the transmission performance of sub-threshold signal and induce various chaotic resonances in coupled neural networks (35).

The state of neurons can be represented by patterns. The pattern no longer looks so smooth when the brain stores short-term memory. There is synchrony in the activity of neurons. In pattern dynamics, synchronization can be induced by Turing instability. Scholars have built various mathematical models and analyzed neurons using Turing dynamic theory to understand the mechanism of memory formation. Zheng et al. studied the effect of noise on the bistable state of the FHN model and explained the biological mechanism of short-term memory by the pattern dynamics theory (36). They also studied the conditions of Turing pattern generation in the Hindmarsh-Rose (HR) model and found that collected current and outgoing current greatly influenced neuronal activity and used this to explain the mechanism of short-term memory generation (37). Wang and Shi proposed the time-delay memristive HR neuron model, found multiple modes and coherence resonance, and speculated that it might be related to the memory effect of neurons (38). We study the FHN model under a multilayer network to get closer to the actual brain structure. The biological mechanism of short-term memory generation is explained by the pattern characteristics of the model. The article is structured as follows. In the next section, firstly, the stability of the equilibrium point in the FHN model is analyzed. Then the sufficient conditions for the Turing instability of the FHN model on the Cartesian product network are found using the comparison principle. Finally, the properties of periodic solutions in FHN multilayer networks are studied using the center manifold theorem. Explaining the mechanism of short-term memory by numerical simulation in Section 3.

2. Description of the FHN model

We consider the general FHN model

$$\begin{aligned}\frac{du}{dt} &= c(u - u^3/3 - av + I),\\ \frac{dv}{dt} &= c(bu - v + d), \end{aligned} \tag{1}$$

Where u is membrane potential, which is a fast variable, and v is recovery variable, which is a slow variable. I is the external input current. a, b represent respectively the intensity of action from v to u and from u to v. And the parameters $c \neq 0$, d are constants. The equilibrium point of the system (Equation 1) satisfies $u^3 + 3(ab-1)u + 3(ad-I) = 0$. Therefore, we have the following conclusion.

Lemma 1 Let $\varpi = \frac{-1+\imath\sqrt{3}}{2}$ and $\varrho = \frac{3}{2}(ad-I)$, in which $\imath$ is the imaginary unit. The influence of parameters on the number of equilibrium of the system (Equation 1) is as follows.

(i) When $ab-1 = ad-I = 0$, the equation has triple zero roots and the trace of the system (1) at that point is constant 0.

(ii) When $\Delta = \varrho^2 + (ab-1)^3 > 0$, the equation has only one real root $\sqrt[3]{-\varrho+\sqrt{\Delta}} + \sqrt[3]{-\varrho-\sqrt{\Delta}}$.

(iii) When $\Delta = 0$, $ab \neq 1$ and $ad \neq I$, the equation has two real roots $-2\sqrt[3]{\varrho}$ and $\sqrt[3]{\varrho}$, and the determinant at the second root $\sqrt[3]{\varrho}$ of system (1) is always 0.

(iv) When $\Delta < 0$, the equation has three unequal real roots $\sqrt[3]{-\varrho+\sqrt{\Delta}} + \sqrt[3]{-\varrho-\sqrt{\Delta}}$, $\varpi\sqrt[3]{-\varrho+\sqrt{\Delta}} + \varpi^2\sqrt[3]{-\varrho-\sqrt{\Delta}}$, $\varpi^2\sqrt[3]{-\varrho+\sqrt{\Delta}} + \varpi\sqrt[3]{-\varrho-\sqrt{\Delta}}$.

Let $U^* = (u^*, v^*)$ be the equilibrium point of the system (1). By coordinate transformation $\bar{u} = u(t) - u^*$, $\bar{v} = v(t) - v^*$, we get the following equivalent system. For convenience, $u(t), v(t)$ are still used to denote $\bar{u}, \bar{v}$,

$$\begin{aligned} \frac{du}{dt} &= a_{11}u + a_{12}v + f(u), \\ \frac{dv}{dt} &= a_{21}u + a_{22}v, \end{aligned} \tag{2}$$

where $a_{11} = c(1 - u^{*^2})$, $a_{12} = -ac$, $a_{21} = bc$, $a_{22} = -c$, $f(u) = -u^*u^2 - \frac{u^3}{3}$. The corresponding determinant $\Delta_0 = c^2(u^{*^2} + ab - 1)$ and the trace $Tr_0 = -cu^{*^2}$. By the Routh-Hurwitz criterion, the equilibrium $(0, 0)$ of the system (Equation 2) is stable if and only if (H 1) holds,

$$cu^{*^2} > 0 \text{ and } u^{*^2} + ab - 1 > 0. \tag{H 1}$$

2.1. FHN model on Cartesian product network

Now we discuss the effect of the Cartesian product networks on the stability of the equilibrium point $(0, 0)$. Two networks R and E with n_r and n_e nodes are given, respectively. $(L^R) = A^R - (k_i\delta_{ij})^R$ $((L^E) = A^E - (k_i\delta_{ij})^E)$ is the Laplacian matrix of the network R (E). A is the adjacency matrix of the network. And k_i denotes the degree of the ith node. δ_{ij} satisfies, $\delta_{ij} = 1$ when node i has an edge with node j; $\delta_{ij} = 0$ when there is no edge. By using the Kronecker product, we can get the Cartesian product network $R\square E$ ($\square$ stands for multilayer network), which has $n_r n_e$ nodes. Then the Laplacian matrix of $R\square E$ is denoted as

$$L^{R\square E} = L^R \otimes \mathbb{I}_{n_e} + \mathbb{I}_{n_r} \otimes L^E,$$

and the eigenvalues of $R\square E$ are of form

$$\Lambda_{\alpha\beta}^{R\square E} = \Lambda_\alpha^R + \Lambda_\beta^E, \quad \alpha \in \{1, \cdots, n_r\}, \quad \beta \in \{1, \cdots, n_e\}.$$

A general FHN model on Cartesian product network can be expressed as

$$\begin{aligned} \frac{du_{re}}{dt} &= a_{11}u_{re} + a_{12}v_{re} + f(u_{re}) + \mathrm{L}_u u_{re}, \\ \frac{dv_{re}}{dt} &= a_{21}u_{re} + a_{22}v_{re} + \mathrm{L}_v v_{re}, \end{aligned} \tag{3}$$

Where $r \in \{1, \cdots, n_r\}$, $e \in \{1, \cdots, n_e\}$. The Laplacian operator L_u is

$$\mathrm{L}_u = D_u^R L^R \otimes \mathbb{I}_{n_e} + D_u^E \mathbb{I}_{n_r} \otimes L^E.$$

D_u^R, D_v^R (D_u^E, D_v^E) are the diffusion coefficients of the network R (E). Notice that $(L^R \otimes \mathbb{I}_{n_e})(u_{re}) = (L^R u_r^R, u_e^E) = \sum_{r'} L_{rr'}^R u_{r'e}$, and similarly, we can get $\mathbb{I}_{n_r} \otimes L^E$. For $\mathrm{L}_v v_{re}$, we can get similar result. Expanding u_{re} and v_{re} in Fourier space, we can obtain linearized equation for equation (3),

$$\begin{aligned} \frac{du_{re}}{dt} &= a_{11}u_{re} + a_{12}v_{re} + (D_u^R\Lambda_\alpha^R + D_u^E\Lambda_\beta^E)u_{re}, \\ \frac{dv_{re}}{dt} &= a_{21}u_{re} + a_{22}v_{re} + (D_v^R\Lambda_\alpha^R + D_v^E\Lambda_\beta^E)v_{re}. \end{aligned} \tag{4}$$

Lemma 2 Comparison principles Consider the ODE

$$\frac{d^2S}{dt} + P(t)\frac{dS}{dt} + Q(t)S = 0, \tag{A 1}$$

and suppose that there exists some $\Phi(t)$ such that

$$Q(t) \leq -\frac{1}{\Phi}\frac{d^2\Phi}{dt} - \frac{1}{\Phi}\frac{d\Phi}{dt}P(t), \quad \forall t \in \Omega. \tag{A 2}$$

If (A 2) holds, then the fundamental solution $S(t)$ of (A 1) satisfies $|S| \geq \Phi(t)$ for all $t \in \Omega$. In particular, $S(t)$ has an exponential growth rate on Ω if $Q(t) < 0$ for all $t \in \Omega$.

The proof of the lemma is divided into two cases. Let's discuss it first at the boundary, and then prove it on the inside by using the properties of the Riccati equation. The detailed proof can be seen in Van Gorder (39).

Theorem 1 Assume that (H 1) holds.

$$\begin{aligned} &\Delta_0 + \Lambda_\beta^E(a_{22}D_u^E + a_{11}D_v^E) + \Lambda_\alpha^R(a_{22}D_u^R + a_{11}D_v^R) \\ &+ (\Lambda_\beta^E)^2 D_u^E D_v^E + \Lambda_\beta^E\Lambda_\alpha^R(D_u^E D_v^R + D_u^R D_v^E) + (\Lambda_\alpha^R)^2 D_u^R D_v^R < 0. \end{aligned} \tag{H 2}$$

If (H 2) holds, then $(0, 0)$ for the system (Equation 3) is linearly unstable.

Proof We consider the Equation (4). Separating v_{re} from the first equation of Equation (4), we can obtain

$$v_{re} = \frac{u_{re'} - a_{11}u_{re} - (D_u^R\Lambda_\alpha^R + D_u^E\Lambda_\beta^E)u_{re}}{a_{12}}.$$

Putting it into the second equation of Equation (4), we can obtain a second-order ODE about u_{re},

$$\begin{aligned} u_{re''} &- [Tr_0 + \Lambda_\beta^E(D_u^E + D_v^E) + \Lambda_\alpha^R(D_u^R + D_v^R)]u_{re'} \\ &+ [\Delta_0 + \Lambda_\beta^E(a_{22}D_u^E + a_{11}D_v^E) + \Lambda_\alpha^R(a_{22}D_u^R + a_{11}D_v^R) \\ &+ (\Lambda_\beta^E)^2 D_u^E D_v^E + \Lambda_\beta^E\Lambda_\alpha^R(D_u^E D_v^R + D_u^R D_v^E) \\ &+ (\Lambda_\alpha^R)^2 D_u^R D_v^R]u_{re} = 0. \end{aligned}$$

Similarly, we get a second-order ODE about v_{re}.

According to Lemma 2, a sufficient condition (H 2) for Turing instability caused by the Cartesian product network at $(0, 0)$ is obtained. Of course, networks do not always cause instability.

2.2. The Hopf bifurcation of FHN network caused by delay

Suppose that $(0, 0)$ in Equation (3) is stable, we next consider the effect of time delay on $(0, 0)$. Adding time delay to the FHN network model (Equation 3), we have

$$\begin{aligned} \frac{du_{re}}{dt} &= a_{11}u_{re} + a_{12}v_{re}(t - \tau) + f(u_{re}) + \mathrm{L}_u u_{re}, \\ \frac{dv_{re}}{dt} &= a_{21}u_{re} + a_{22}v_{re} + \mathrm{L}_v v_{re}. \end{aligned} \tag{5}$$

The Jacobian matrix of each node becomes

$$\begin{aligned} J_{re} = &\begin{pmatrix} a_{11} + D_u^R\Lambda_\alpha^R + D_u^E\Lambda_\beta^E & 0 \\ a_{21} & a_{22} + D_v^R\Lambda_\alpha^R + D_v^E\Lambda_\beta^E \end{pmatrix} \\ &+ \begin{pmatrix} 0 & a_{12} \\ 0 & 0 \end{pmatrix} \mathrm{e}^{-\lambda_{re}\tau} \triangleq J^0 + J^1\mathrm{e}^{-\lambda_{re}\tau}. \end{aligned}$$

Then the transcendental equation of the system (Equation 5) at $(0, 0)$ is

$$\lambda_{re}^2 + B_1\lambda_{re} + B_2 + B_3 \mathrm{e}^{-\lambda_{re}\tau} = 0,$$

where

$$\begin{aligned} B_1 &= -Tr_0 - (D_u^E + D_v^E)\Lambda_\beta^E - (D_u^R + D_v^R)\Lambda_\alpha^R, \\ B_2 &= (D_v^E\Lambda_\beta^E + D_v^R\Lambda_\alpha^R + a_{22})(D_u^E\Lambda_\beta^E + D_u^R\Lambda_\alpha^R + a_{11}), \\ B_3 &= -a_{12}a_{21}. \end{aligned}$$

Suppose $\imath\omega$ $(\omega > 0)$ be a root of the transcendental equation. And substituting $\imath\omega$ into the above equation, we can obtain

$$-\omega^2 + B_2 + B_3 cos(\omega\tau) + \imath\big(B_1\omega - B_3 sin(\omega\tau)\big) = 0.$$

Comparing the coefficients, we have

$$\begin{cases} B_3 cos(\omega\tau) = \omega^2 - B_2, \\ B_3 sin(\omega\tau) = B_1\omega, \end{cases}$$

then we obtain

$$\omega^4 + (B_1^2 - 2B_2)\omega^2 + B_2^2 - B_3^2 = 0.$$

Let $x = \omega^2$, $p = B_1^2 - 2B_2$, $q = B_2^2 - B_3^2$, then the equation becomes

$$x^2 + px + q = 0. \tag{6}$$

Lemma 3 Assume that $(0, 0)$ in Equation (3) is stable. If $4q < 0 \leq p^2$ and $p > 0$, then the real parts of all roots of the transcendental equation are less than 0 for $\tau \in [0, \tau_0)$ and $\frac{dRe\lambda_{re}(\tau_0)}{d\tau} \neq 0$.
Proof The Equation (6) has only one positive root when $4q < 0 \leq p^2$ and $p > 0$, denoted by x_0. Hence, $\imath\omega_0 = \imath\sqrt{x_0}$ is a purely imaginary root of the transcendental equation. Let

$$\tau_0^j(\Lambda_\alpha^R, \Lambda_\beta^E) = \frac{1}{\omega_0} \arccos \frac{\omega_0^2 - B_2}{B_3} + 2\pi j, \qquad j = 0, 1, 2, \cdots.$$

Define

$$\tau_0 = \min_{j \geq 1} \tau_0^j(\Lambda_\alpha^R, \Lambda_\beta^E). \tag{7}$$

Then again, τ_0 is the minimum value of τ_0^j, so the real parts of all roots of the transcendental equation are less than 0 for $\tau \in [0, \tau_0)$.

Next we prove the transversal condition. Let

$$\lambda_{re}(\tau) = \eta(\tau) + \imath\omega(\tau)$$

be the root of transcendental equation, then $\eta(\tau_0) = 0$, $\omega(\tau_0) = \omega_0$. By taking the derivative with respect to τ in the transcendental equation, we can get

$$\frac{d\lambda_{re}(\tau)}{d\tau} = \frac{B_3\lambda_{re}\mathrm{e}^{-\lambda_{re}\tau}}{2\lambda_{re} + B_1 - B_3\tau\mathrm{e}^{-\lambda_{re}\tau}}.$$

Substituting ω_0, τ_0 into the above equation, we can obtain

$$\frac{dRe\lambda_{re}(\tau_0)}{d\tau} = \frac{\omega_0^2}{\Theta}(2\omega_0^2 + p),$$

where

$$\Theta = (-\omega_0^2\tau_0 + B_2\tau_0 + B_1)^2 + (B_1\omega_0\tau_0 + 2\omega_0)^2.$$

So

$$\frac{dRe\lambda_{re}(\tau_0)}{d\tau} \neq 0.$$

According to the above analysis, the system (Equation 5) will occur Hopf bifurcation at $\tau = \tau_0$ when Lemma 3 holds. Next, we discuss the properties of periodic solutions. The idea is: firstly, the system is written in the form of abstract ODE by using the infinitesimal generators theorem; then, A two-dimensional ODE that is the restriction to its center manifold is obtained by using the spectral decomposition theorem and the central manifold theorem of infinite dimensional systems; finally, the Hassard method is applied to determine the bifurcation attributes' parameters. The delay τ is taken as the control parameter, and let $\tau = \tau_0 + \grave{o}$, $t = \tau\varsigma$. For convenience, we'll still use t to stand for ς. Setting $\aleph(t) = (u_{re}(t), v_{re}(t))^T$ be the solution of system (Equation 5) and define $\aleph_t(\theta) = \aleph(t + \theta)$, $\theta \in [-1, 0]$.

The system (Equation 5) is transformed into the following functional equation,

$$\dot{\aleph}_t = \mathrm{AE}_{\grave{o}}\aleph_t + F_{\grave{o}}(\aleph_t), \tag{8}$$

where linear operator $\mathrm{AE}_{\acute{o}} : C([-1, 0], \mathbb{R}^2) \triangleq C \to \mathbb{R}^2$,

$$\mathrm{AE}_{\acute{o}}\phi = (\tau_0 + \grave{o})J^0\phi(0) + (\tau_0 + \grave{o})J^1\phi(-1);$$

nonlinear operator $F_{\grave{o}} : C \to \mathbb{R}^2$,

$$F_{\grave{o}}(\phi) = (\tau_0 + \grave{o}) \begin{pmatrix} -u^*\phi_1(0)^2 - \frac{\phi_1(0)^3}{3} \\ 0 \end{pmatrix},$$

where $\phi(\theta) = (\phi_1(\theta), \phi_2(\theta))^T$.

From Riesz representation theorem, there exists matrix $\eta(\theta, \grave{o})$ of bounded variation functions satisfying

$$\mathrm{AE}_{\grave{o}}\phi = \int_{-1}^{0} \phi(\theta)\mathrm{d}\eta(\theta, \grave{o}), \quad \text{where } \phi \in C.$$

Let

$$\eta(\theta, \grave{o}) = (\tau_0 + \grave{o})J^0\delta(\theta) + (\tau_0 + \grave{o})J^1\delta(\theta + 1),$$

where $\delta(\cdot)$ denotes Dirac function. According to the infinitesimal generators theorem, the abstract differential equation can be obtained from Equation (8)

$$\dot{\aleph}_t = A_{\grave{o}}\aleph_t + R_{\grave{o}}(\aleph_t), \tag{9}$$

where

$$A_{\grave{o}}\phi(\theta) = \begin{cases} \frac{\mathrm{d}\phi(\theta)}{\mathrm{d}\theta}, & \theta \in [-1, 0), \\ \int_{-1}^{0} \mathrm{d}_\sigma \eta(\grave{o}, \sigma)\phi(\sigma), & \theta = 0; \end{cases}$$

$$R_{\grave{o}}(\phi(\theta)) = \begin{cases} 0, & \theta \in [-1, 0), \\ F_{\grave{o}}(\phi), & \theta = 0. \end{cases}$$

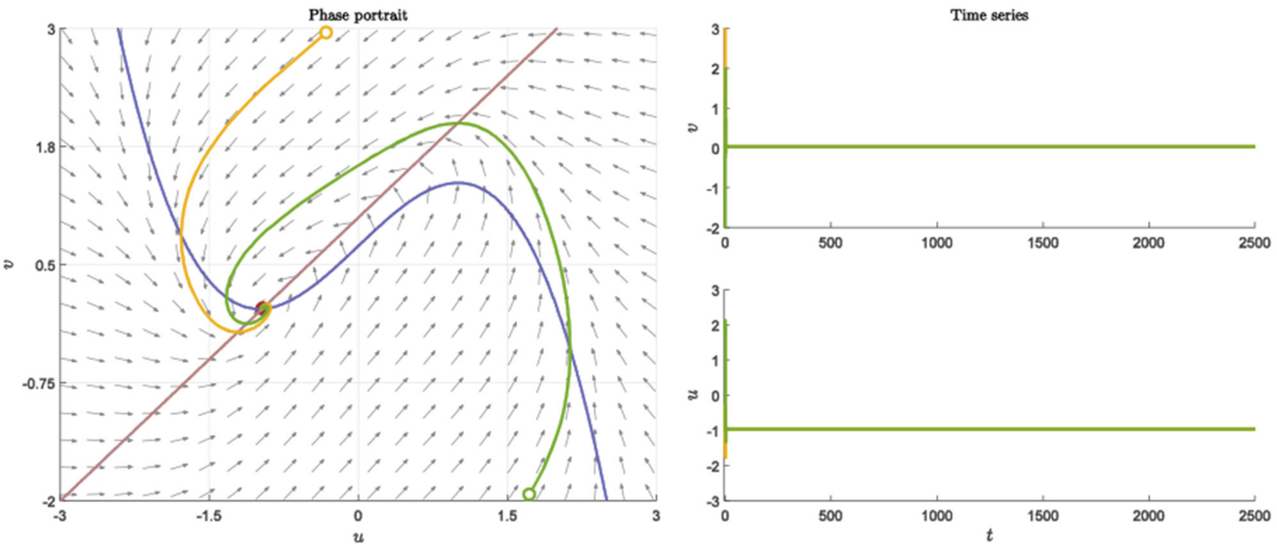


FIGURE 1
Nullcline, phase portraits with different initial value and time series when $a = 1,\ b = 1,\ c = 2,\ d = 1,\ I = 0.7$.

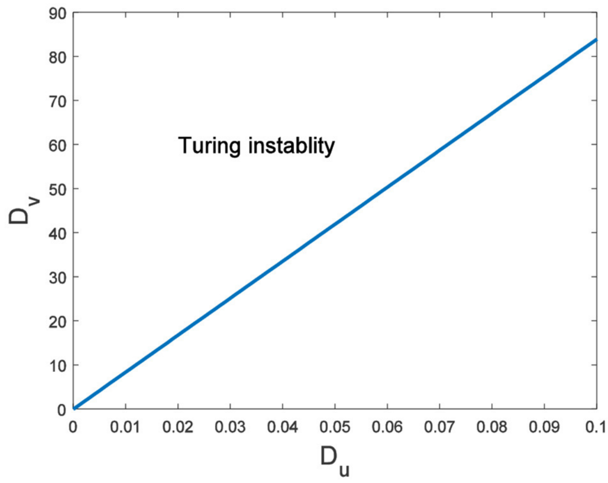


FIGURE 2
The range of Turing instability in the general diffusion system about D_u and D_v.

In the following, we will discuss ODE (Equation 9) by using formal adjoint theorem, center manifold theorem and normal form theory.

Let $A_{\grave{o}}^*$ be the conjugate operator of $A_{\grave{o}}$. According to the formal adjoint theorem, there is

$$A_{\grave{o}}^*\psi(s) = \begin{cases} -\frac{\mathrm{d}\psi(s)}{\mathrm{d}s}, & s \in (0,1], \\ \int_{-1}^{0} \mathrm{d}\eta^T(\sigma,0)\psi(-\sigma), & s = 0. \end{cases}$$

Define product

$$\langle \psi(s), \phi(\theta)\rangle = \bar{\psi}^T(0)\phi(0) - \int_{\theta=-1}^{0}\int_{\xi=0}^{\theta} \bar{\psi}^T(\xi-\theta)\mathrm{d}\eta(\theta)\phi(\xi)\mathrm{d}\xi,$$

which satisfies $\langle \psi, A_{\grave{o}}\phi\rangle = \langle A_{\grave{o}}^*\psi, \phi\rangle$ and $\eta(\theta) = \eta(\theta, 0)$. From the previous discussion, we can obtain that $\pm\imath\omega_0\tau_0$ are the eigenvalues of A_0, A_0^*.

Lemma 4 Let $q(\theta) = (1, q_2)^T \mathrm{e}^{\imath\omega_0\tau_0\theta}$ be the eigenvector of A_0 corresponding to $\imath\omega_0\tau_0$, and $q^*(s) = \kappa(q_1^*, 1)^T\mathrm{e}^{\imath\omega_0\tau_0 s}$ be the eigenvector of A_0^* corresponding to $-\imath\omega_0\tau_0$. And let $\langle q^*(s), q(\theta)\rangle = 1$, then we can choose

$$q_2 = \frac{a_{21}}{\imath\omega_0 - a_{22}}, \quad q_1^* = \frac{-a_{21}}{\imath\omega_0 + a_{11}},$$
$$\kappa = \frac{1}{q_1^* + \bar{q_2} + \tau_0 a_{12} q_1^* \bar{q_2}\mathrm{e}^{\imath\omega_0\tau_0}}.$$

Proof From the hypothesis, we have

$$A_0\begin{pmatrix}1\\ q_2\end{pmatrix} = \imath\omega_0\tau_0\begin{pmatrix}1\\ q_2\end{pmatrix}, \quad A_0^*\begin{pmatrix}q_1^*\\ 1\end{pmatrix} = -\imath\omega_0\tau_0\begin{pmatrix}q_1^*\\ 1\end{pmatrix}.$$

Then

$$q_2 = \frac{a_{21}}{\imath\omega_0 - a_{22}}, \quad q_1^* = \frac{-a_{21}}{\imath\omega_0 + a_{11}}.$$

Next, we calculate the expression of κ. According to the bilinear inner product formula, we have

$$\begin{aligned}\langle q^*(s), q(\theta)\rangle &= \bar{q^*}^T(0)q(0) - \int_{-1}^{0}\int_{0}^{\theta} \bar{q^*}^T(\xi-\theta)\mathrm{d}\eta(\theta)q(\xi)\mathrm{d}\xi \\ &= \bar{\kappa}(\bar{q_1^*}, 1)(1, q_2)^T \\ &- \int_{-1}^{0}\int_{0}^{\theta} \bar{\kappa}(\bar{q_1^*}, 1)\mathrm{e}^{-\imath\omega_0\tau_0(\xi-\theta)}\mathrm{d}\eta(\theta)(1, q_2)^T\mathrm{e}^{\imath\omega_0\tau_0\xi}\mathrm{d}\xi \\ &= \bar{\kappa}\big[\bar{q_1^*} + q_2 - (\bar{q_1^*}, 1)\int_{-1}^{0}\theta\mathrm{e}^{\imath\omega_0\tau_0\theta}\mathrm{d}\eta(\theta)(1, q_2)^T\big] \\ &= \bar{\kappa}\big[\bar{q_1^*} + q_2 + (\bar{q_1^*}, 1)\tau_0 J^1\mathrm{e}^{-\imath\omega_0\tau_0}(1, q_2)^T\big] \\ &= \bar{\kappa}\big[\bar{q_1^*} + q_2 + \tau_0 a_{12}\bar{q_1^*}q_2\mathrm{e}^{-\imath\omega_0\tau_0}\big].\end{aligned}$$

To make $\langle q^*(s), q(\theta)\rangle = 1$, we take

$$\bar{\kappa} = \frac{1}{\bar{q}_1^* + q_2 + \tau_0 a_{12}\bar{q}_1^* q_2 e^{-\imath\omega_0\tau_0}}.$$

The center manifold Ω_0 of Equation (8) is locally invariant when $\grave{o} = 0$. To achieve spectral decomposition, we build the local coordinates z and $\bar{z}$ on the center manifold Ω_0. Let $U_t = U_t(\theta)$ be the solution of the system (Equation 8) when $\grave{o} = 0$, then

$$z(t) = \langle q^*, U_t\rangle.$$

And let

$$W(t,\theta) = U_t(\theta) - z(t)q(\theta) - \bar{z}(t)\bar{q}(\theta). \tag{10}$$

$W(t,\theta) = W(z,\bar{z},\theta)$ on the center manifold Ω_0, so $W(z,\bar{z},\theta)$ can be expanded as

$$W(z,\bar{z},\theta) = W_{20}(\theta)\tfrac{z^2}{2} + W_{11}(\theta)z\bar{z} + W_{02}(\theta)\tfrac{\bar{z}^2}{2} + \cdots. \tag{11}$$

W is real when U_t is real. Therefore, in this case, let's just look at the real solution. Obviously, there is

$$\langle q^*, W\rangle = 0.$$

Because of the existence of the center manifold, it is possible to transform the functional differential equation (Equation 8) into simple complex variable ODE on Ω. When $\grave{o} = 0$, there is

$$\begin{aligned}\dot{z}(t) &= \langle q^*, \dot{U}_t\rangle \\ &= \imath\omega_0\tau_0 z(t) + \bar{q^*}^T(0)F_0\big(W(z,\bar{z},\theta) + zq(\theta) + \bar{z}\bar{q}(\theta)\big) \\ &\triangleq \imath\omega_0\tau_0 z(t) + g(z,\bar{z})(t).\end{aligned} \tag{12}$$

And since $F_{\grave{o}}(\phi)$ is at least quadratic with respect to ϕ, we can write

$$g(z,\bar{z}) = g_{20}\tfrac{z^2}{2} + g_{11}z\bar{z} + g_{02}\tfrac{\bar{z}^2}{2} + g_{21}\tfrac{z^2\bar{z}}{2} + \cdots. \tag{13}$$

By combining Equations (10), (11), we can obtain

$$\begin{aligned}U_t(\theta) &= W(t,\theta) + z(t)q(\theta) + \bar{z}(t)\bar{q}(\theta) \\ &= (1,q_2)^T e^{\imath\omega_0\tau_0\theta}z + (1,\bar{q}_2)^T e^{-\imath\omega_0\tau_0\theta}\bar{z} + W_{20}(\theta)\frac{z^2}{2} \\ &\quad + W_{11}(\theta)z\bar{z} + W_{02}(\theta)\frac{\bar{z}^2}{2} + \cdots.\end{aligned}$$

Substituting the above equation into (13), it can be obtained

$$\begin{aligned}&g(z,\bar{z}) = \bar{q^*}^T(0)F_0(U_t) = \tau_0\bar{\kappa}(\bar{q}_1^*, 1)\begin{pmatrix}-u^*\phi_1(0)^2 - \frac{\phi_1(0)^3}{3} \\ 0\end{pmatrix} \\ &= -\tau_0\bar{\kappa}\bar{q}_1^*\big[u^*(z + \bar{z} + W_{20}^{(1)}(0)\frac{z^2}{2} + W_{11}^{(1)}(0)z\bar{z} + W_{02}^{(1)}(0)\frac{\bar{z}^2}{2} + \cdots)^2 \\ &+ \frac{1}{3}(z + \bar{z} + W_{20}^{(1)}(0)\frac{z^2}{2} + W_{11}^{(1)}(0)z\bar{z} + W_{02}^{(1)}(0)\frac{\bar{z}^2}{2} + \cdots)^3\big] \\ &= -\tau_0\bar{\kappa}\bar{q}_1^*\big[u^*\bar{z}^2 + 2u^*z\bar{z} + u^*z^2 \\ &+ (2u^*W_{11}^{(1)}(0) + u^*W_{20}^{(1)}(0) + 1)z^2\bar{z} + \cdots\big].\end{aligned}$$

Obviously, there are

$$\begin{aligned}g_{02} &= g_{11} = g_{20} = -2\tau_0\bar{\kappa}\bar{q}_1^*u^*, \\ g_{21} &= -2\tau_0\bar{\kappa}\bar{q}_1^*\big(2u^*W_{11}^{(1)}(0) + u^*W_{20}^{(1)}(0) + 1\big).\end{aligned} \tag{14}$$

Observing the above equation, we can see that if we want to calculate g_{21}, we must first calculate $W_{20}(\theta)$ and $W_{11}(\theta)$. Next, we determine the exact expression for $W_{20}(\theta)$, $W_{11}(\theta)$.

According to Equations (9), (10), and (12), we have

$$\begin{aligned}\dot{W} &= \dot{U}_t - \dot{z}q - \dot{\bar{z}}\bar{q} \\ &= \begin{cases}A_0W - gq(\theta) - \bar{g}\bar{q}(\theta), & -1 \le \theta < 0 \\ A_0W - gq(\theta) - \bar{g}\bar{q}(\theta) + F_0, & \theta = 0\end{cases} \\ &\triangleq A_0W + M(z,\bar{z},\theta),\end{aligned} \tag{15}$$

where

$$M(z,\bar{z},\theta) = M_{20}(\theta)\tfrac{z^2}{2} + M_{11}(\theta)z\bar{z} + M_{02}(\theta)\tfrac{\bar{z}^2}{2} + \cdots. \tag{16}$$

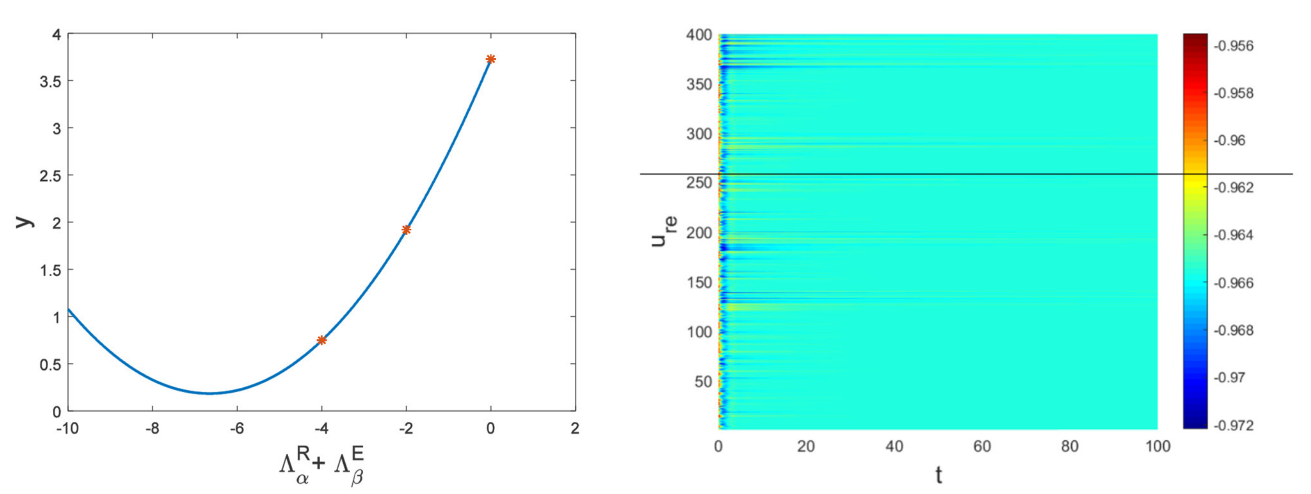


FIGURE 3
The relationship between $\Lambda_\alpha^R + \Lambda_\beta^E$ and y and the corresponding pattern when $a = 1,\ b = 1,\ c = 2,\ d = 1,\ l = 0.7,\ D_u = 0.01,\ D_v = 8,\ p = 0.01$. The red dots are the eigenvalues of the network Laplacian matrix.

Combining Equations (11), (15), and (16), $\dot{W}$ can be expressed as

$$\begin{aligned}\dot{W} &= A_0\Big[W_{20}(\theta)\frac{z^2}{2} + W_{11}(\theta)z\bar{z} + W_{02}(\theta)\frac{\bar{z}^2}{2} + \cdots\Big] \\ &\quad + M_{20}(\theta)\frac{z^2}{2} + M_{11}(\theta)z\bar{z} + M_{02}(\theta)\frac{\bar{z}^2}{2} + \cdots \\ &= \Big[A_0W_{20}(\theta) + M_{20}(\theta)\Big]\frac{z^2}{2} + \Big[A_0W_{11}(\theta) + M_{11}(\theta)\Big]z\bar{z} \\ &\quad + \Big[A_0W_{02}(\theta) + M_{02}(\theta)\Big]\frac{\bar{z}^2}{2} + \cdots.\end{aligned} \tag{17}$$

On the other hand, combining Equations (11), (12), we know that on the center manifold Ω_0 near the origin, $\dot{W}$ can also be expressed as

$$\begin{aligned}\dot{W} &= (W_{20}z + W_{11}\bar{z} + \cdots)\big[\imath\omega_0\tau_0 z + g(z,\bar{z})\big] \\ &\quad + (W_{11}z + W_{02}\bar{z} + \cdots)\big[-\imath\omega_0\tau_0\bar{z} + \bar{g}(z,\bar{z})\big] \\ &= 2\imath\omega_0\tau_0(W_{20}\frac{z^2}{2} + W_{02}\frac{\bar{z}^2}{2} + \cdots).\end{aligned} \tag{18}$$

Comparing the coefficients of z^2 and $z\bar{z}$ in Equations (17), (18), the relationship between $W_{ij}(\theta)$ and $M_{ij}(\theta)$ can be obtained

$$(2\imath\omega_0\tau_0\mathbb{I} - A_0)W_{20}(\theta) = M_{20}(\theta), \quad -A_0W_{11}(\theta) = M_{11}(\theta). \tag{19}$$

Next, we will determine $W_{11}(\theta)$ and $W_{20}(\theta)$ according to the relationship between $g(z,\bar{z})$ and $M(z,\bar{z},\theta)$.

When $-1 \le \theta < 0$, combining Equations (15), (16), it is clear that

$$\begin{aligned}M_{20}(\theta) &= -g_{20}q(\theta) - \bar{g}_{02}\bar{q}(\theta), \\ M_{11}(\theta) &= -g_{11}q(\theta) - \bar{g}_{11}\bar{q}(\theta).\end{aligned} \tag{20}$$

Combining Equations (19), (20) and the definition of $A_{\grave{0}}$, we get

$$\begin{aligned}\frac{\mathrm{d}W_{20}}{\mathrm{d}\theta} &= 2\imath\omega_0\tau_0W_{20}(\theta) + g_{20}q(\theta) + \bar{g}_{02}\bar{q}(\theta), \\ \frac{\mathrm{d}W_{11}}{\mathrm{d}\theta} &= g_{11}q(\theta) + \bar{g}_{11}\bar{q}(\theta).\end{aligned} \tag{21}$$

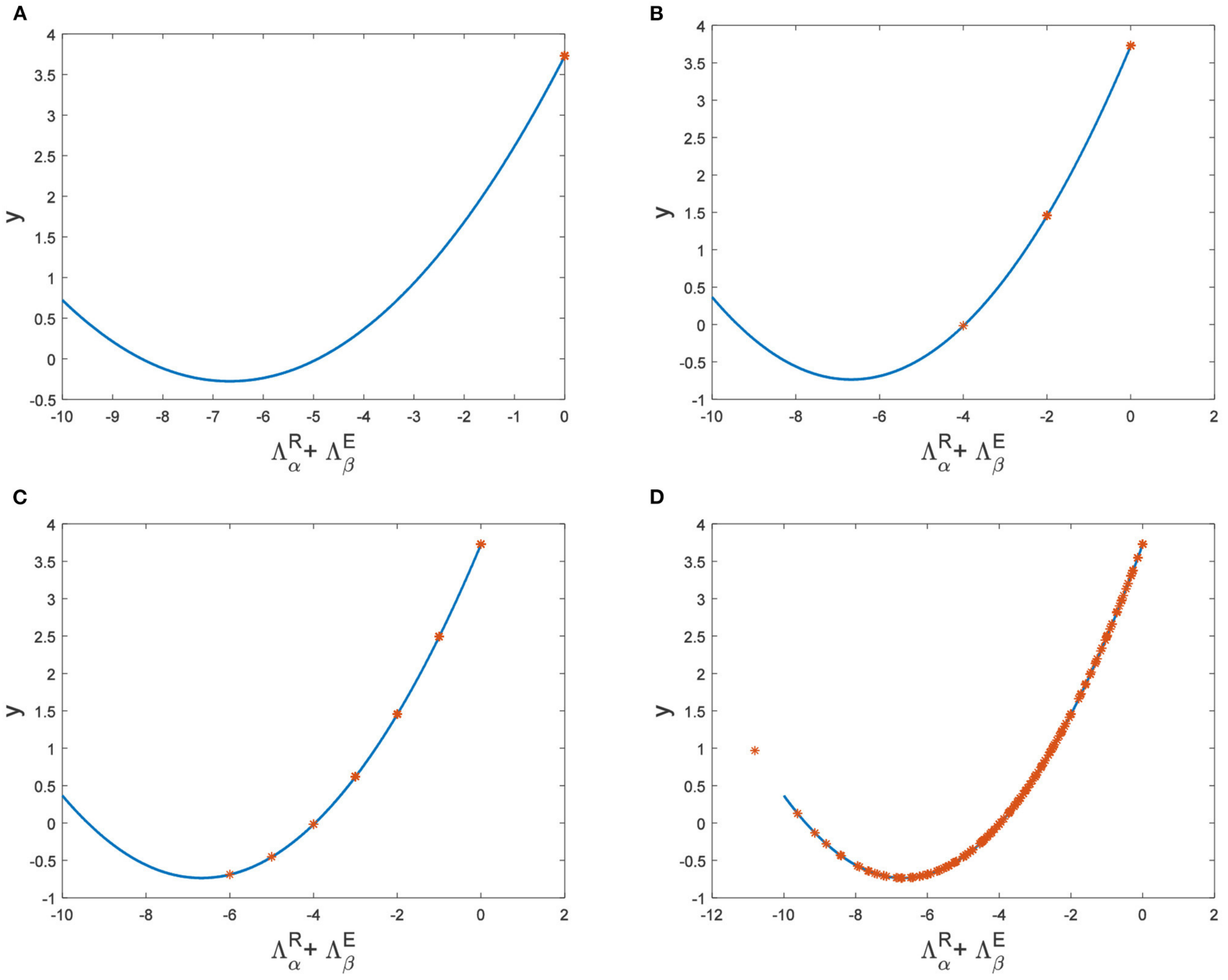


FIGURE 4
The relationship between $\Lambda_\alpha^R + \Lambda_\beta^E$ and y when $a = 1,\ b = 1,\ c = 2,\ d = 1,\ l = 0.7,\ D_u = 0.01$. **(A)** $D_v = 9,\ p = 0.001$. **(B)** $D_v = 10,\ p = 0.006$. **(C)** $D_v = 10,\ p = 0.01$. **(D)** $D_v = 10,\ p = 0.1$.

By substituting $q(\theta) = (1, q_2)^T \mathrm{e}^{\imath\omega_0\tau_0\theta}$ into the above equation, it can be obtained by the constant variation method

$$\begin{aligned} W_{20}(\theta) &= \tfrac{\imath g_{20}}{\omega_0\tau_0} q(0)\mathrm{e}^{\imath\omega_0\tau_0\theta} + \tfrac{\imath\bar{g}_{02}}{3\omega_0\tau_0}\bar{q}(0)\mathrm{e}^{-\imath\omega_0\tau_0\theta} + \ell_1\mathrm{e}^{2\imath\omega_0\tau_0\theta}, \\ W_{11}(\theta) &= -\tfrac{\imath g_{11}}{\omega_0\tau_0} q(0)\mathrm{e}^{\imath\omega_0\tau_0\theta} + \tfrac{\imath\bar{g}_{11}}{\omega_0\tau_0}\bar{q}(0)\mathrm{e}^{-\imath\omega_0\tau_0\theta} + \ell_2, \end{aligned} \tag{22}$$

Where $\ell_1 = (\ell_1^1, \ell_1^2)^T$, $\ell_2 = (\ell_2^1, \ell_2^2)^T$ are two dimensional constant vectors. Next, let's figure out what the values of ℓ_1 and ℓ_2 are.

According to Equation (19) and the definition of A_0, when $\theta = 0$, there is

$$\begin{aligned} &\textstyle\int_{-1}^{0} \mathrm{d}\eta(\theta) W_{20}(\theta) = 2\imath\omega_0\tau_0 W_{20}(0) - M_{20}(0), \\ &\textstyle\int_{-1}^{0} \mathrm{d}\eta(\theta) W_{11}(\theta) = -M_{11}(0). \end{aligned} \tag{23}$$

When $\theta = 0$, combining Equations (15), (16), it is clear that

$$\begin{aligned} M_{20}(0) &= -g_{20}q(0) - \bar{g}_{02}\bar{q}(0) + 2\tau_0 \begin{pmatrix} -u^* \\ 0 \end{pmatrix}, \\ M_{11}(0) &= -g_{11}q(0) - \bar{g}_{11}\bar{q}(0) + 2\tau_0 \begin{pmatrix} -u^* \\ 0 \end{pmatrix}. \end{aligned} \tag{24}$$

Since $q(0)$ is the eigenvector of A_0 corresponding to $\imath\omega_0\tau_0$, we can obtain

$$\begin{aligned} \Big(\imath\omega_0\tau_0\mathbb{I} - \int_{-1}^{0} \mathrm{e}^{\imath\omega_0\tau_0\theta}\mathrm{d}\eta(\theta)\Big)q(0) &= 0, \\ \Big(-\imath\omega_0\tau_0\mathbb{I} - \int_{-1}^{0} \mathrm{e}^{-\imath\omega_0\tau_0\theta}\mathrm{d}\eta(\theta)\Big)\bar{q}(0) &= 0. \end{aligned} \tag{25}$$

Substituting Equations (22), (24), and (25) into Equation (23), we obtain

$$\begin{aligned} &\Big(2\imath\omega_0\tau_0\mathbb{I} - \textstyle\int_{-1}^{0} \mathrm{e}^{2\imath\omega_0\tau_0\theta}\mathrm{d}\eta(\theta)\Big)\ell_1 = 2\tau_0 \begin{pmatrix} -u^* \\ 0 \end{pmatrix}, \\ &\textstyle\int_{-1}^{0} \mathrm{d}\eta(\theta)\ell_2 = -2\tau_0 \begin{pmatrix} -u^* \\ 0 \end{pmatrix}. \end{aligned} \tag{26}$$

When $\grave{o} = 0$, there are

$$\begin{aligned} &\begin{pmatrix} 2\imath\omega_0 - a_{11} - D_u^R\Lambda_\alpha^R - D_u^E\Lambda_\beta^E & -a_{12}\mathrm{e}^{-2\imath\omega_0\tau_0} \\ -a_{21} & 2\imath\omega_0 - a_{22} - D_v^R\Lambda_\alpha^R - D_v^E\Lambda_\beta^E \end{pmatrix}\ell_1 \\ &= 2\begin{pmatrix} -u^* \\ 0 \end{pmatrix}, \\ &\begin{pmatrix} -a_{11} - D_u^R\Lambda_\alpha^R - D_u^E\Lambda_\beta^E & -a_{12} \\ -a_{21} & -a_{22} - D_v^R\Lambda_\alpha^R - D_v^E\Lambda_\beta^E \end{pmatrix}\ell_2 = 2\begin{pmatrix} -u^* \\ 0 \end{pmatrix}, \end{aligned}$$

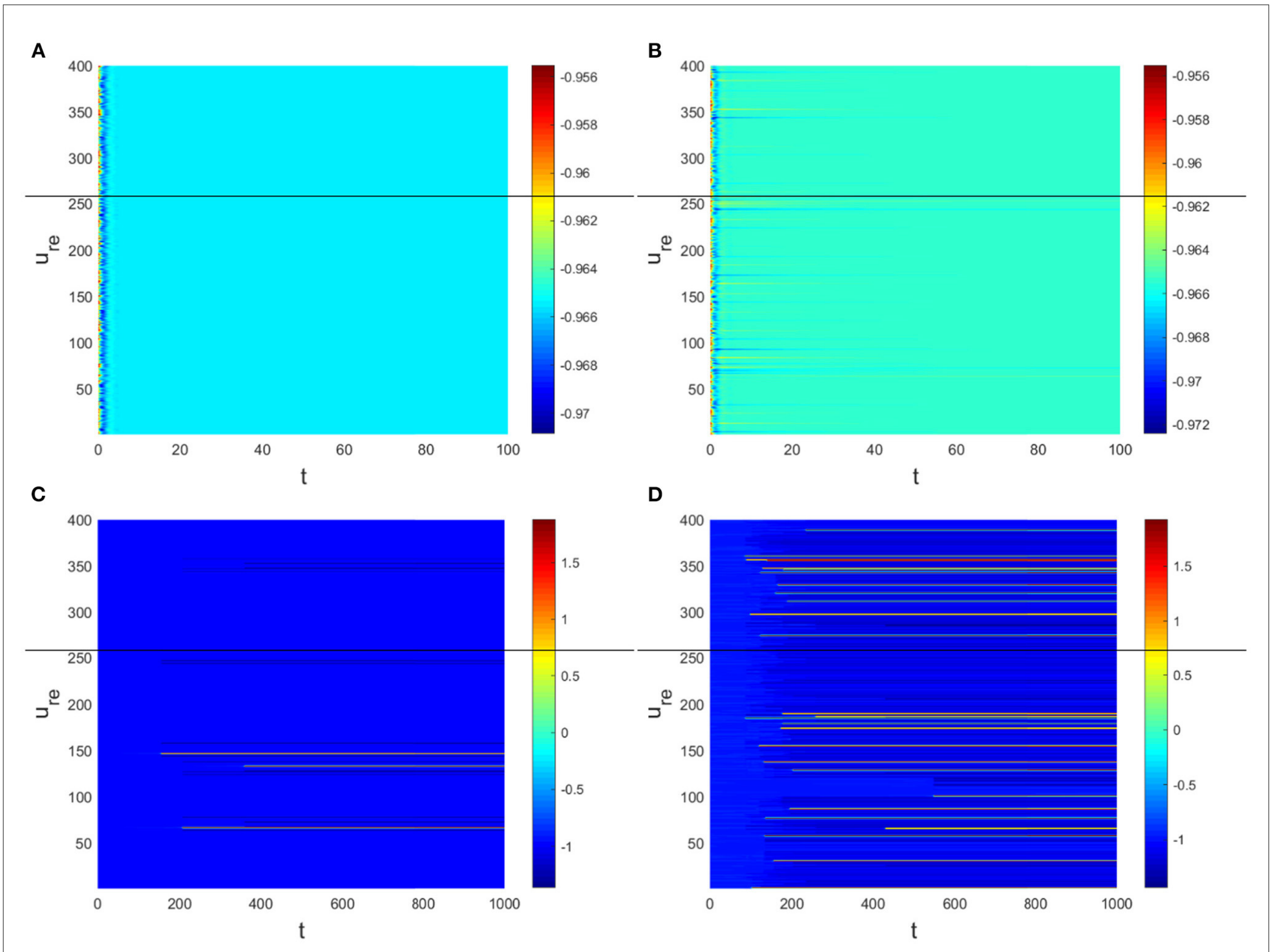


FIGURE 5
The corresponding Turing pattern in Figure 4. **(A)** $D_v = 9,\ p = 0.001$. **(B)** $D_v = 10,\ p = 0.006$. **(C)** $D_v = 10,\ p = 0.01$. **(D)** $D_v = 10,\ p = 0.1$.

that is,

$$\begin{aligned}\ell_1 &= -2\begin{pmatrix} 2\imath\omega_0 - a_{11} - D_u^R\Lambda_\alpha^R - D_u^E\Lambda_\beta^E & -a_{12}e^{-2\imath\omega_0\tau_0} \\ -a_{21} & 2\imath\omega_0 - a_{22} - D_v^R\Lambda_\alpha^R - D_v^E\Lambda_\beta^E \end{pmatrix}^{-1} \\ &\quad\times\begin{pmatrix} u^* \\ 0 \end{pmatrix}, \\ \ell_2 &= -2\begin{pmatrix} -a_{11} - D_u^R\Lambda_\alpha^R - D_u^E\Lambda_\beta^E & -a_{12} \\ -a_{21} & -a_{22} - D_v^R\Lambda_\alpha^R - D_v^E\Lambda_\beta^E \end{pmatrix}^{-1}\begin{pmatrix} u^* \\ 0 \end{pmatrix}.\end{aligned}$$

Substituting ℓ_1, ℓ_2 into Equation (22), we can find W_{20} and W_{11}. To date, g_{20}, g_{21}, g_{11} and g_{02} are now all found, and the normal form Equation (12) that is the restriction to its center manifold is obtained. The key parameters μ_2, T_2 and Floquet exponent β_2 that determine the properties of periodic solutions can be calculated by Hassard's method,

$$\begin{cases} c_1(0) = \frac{\imath}{2\omega_0\tau_0}\left(g_{11}g_{20} - 2|g_{11}|^2 - \frac{1}{3}|g_{02}|^2\right) + \frac{1}{2}g_{21}, \\ \mu_2 = -\frac{Re\left[c_1(0)\right]}{Re\left[\lambda'(\tau_0)\right]}, \\ \beta_2 = 2Re\left[c_1(0)\right], \\ T_2 = -\frac{Im\left[c_1(0)\right]+\mu_2 Im\left[\lambda'(\tau_0)\right]}{\omega_0}. \end{cases} \tag{27}$$

Theorem 2 Suppose that the conditions of Lemma 3 are satisfied, then

(i) If $\mu_2 > 0(< 0)$, the periodic solution is a supercritical (subcritical) Hopf bifurcation.

(ii) If $T_2 > 0(< 0)$, the period of the periodic solution increases (decreases) as τ moves away from τ_0.

(iii) If $\beta_2 > 0(< 0)$, the periodic solutions restricted on the center manifold are orbitally asymptotically unstable (stable).

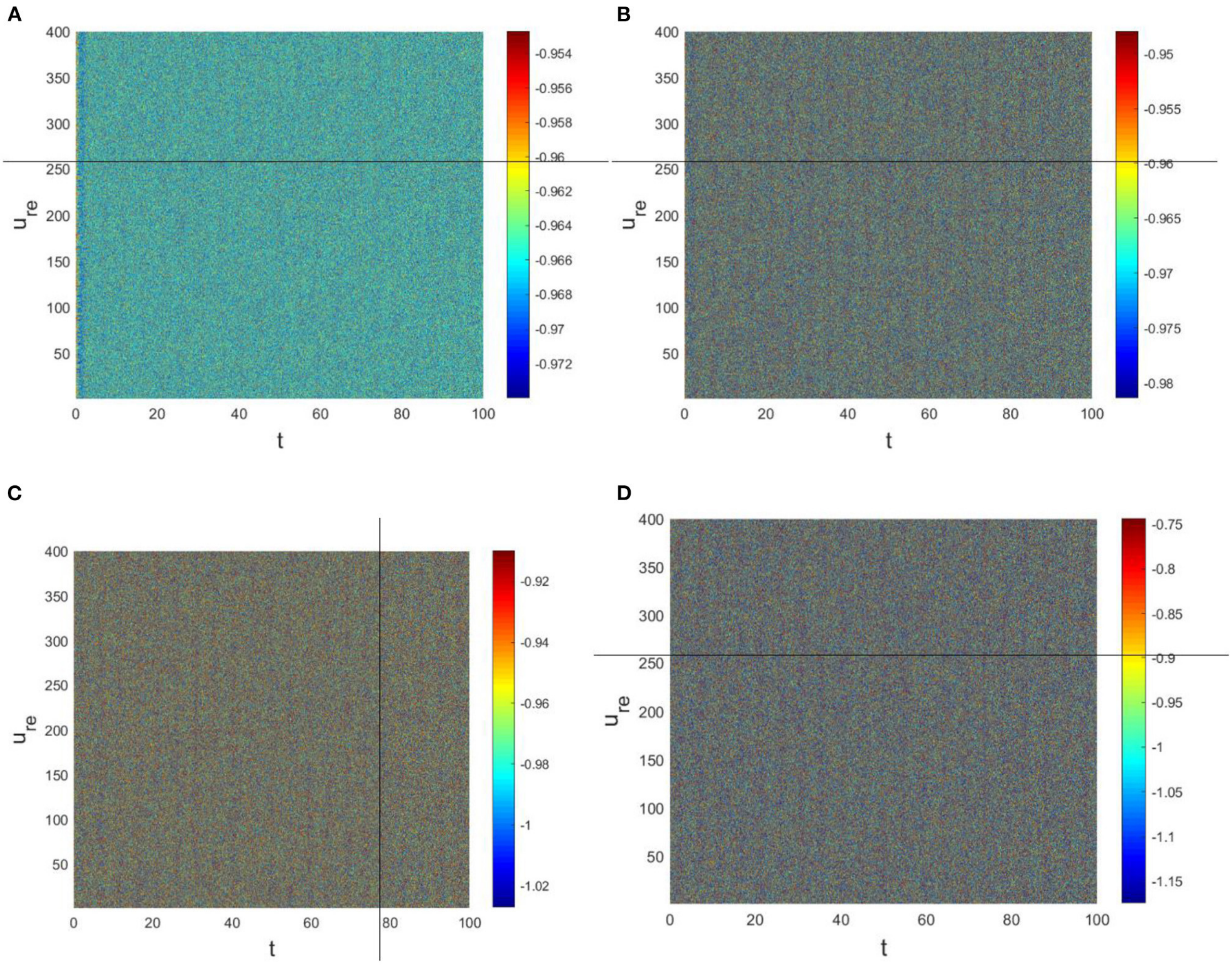


FIGURE 6
Pattern with $a = 1,\ b = 1,\ c = 2,\ d = 1,\ l = 0.7,\ D_u = 0.01,\ D_v = 9,\ p = 0.001$. **(A)** $np = O(10^{-7})$. **(B)** $np = O(10^{-6})$. **(C)** $np = O(10^{-4})$. **(D)** $np = O(10^{-3})$.

3. Simulation

We perform simple simulations to verify the above theoretical results in this section. The topological properties of neural networks are very important to the dynamic behavior of neuronal clusters. In both R and E, we pick random networks with connection probability p and $D_u^R = D_u^E = D_u$, $D_v^R = D_v^E = D_v$. And setting the parameters as $a = 1,\ b = 1,\ c = 2,\ d = 1,\ I = 0.7,\ n_r = n_e = 20$. Neurons still return to the resting state after receiving different stimuli (Figure 1).

In this case, condition (H 2) becomes

$$y = \Delta_0 + (a_{22}D_u + a_{11}D_v)(\Lambda_\alpha^R + \Lambda_\beta^E) + D_uD_v(\Lambda_\alpha^R + \Lambda_\beta^E)^2 < 0.$$

Hence, Turing instability occurs in the general diffusion system when $a_{22}D_u + a_{11}D_v > 0$ and $(a_{22}D_u + a_{11}D_v)^2 - 4\Delta_0 D_uD_v > 0$. And the critical value is $D_v = 8.3923$ when $D_u = 0.01$ (Figure 2). Different dynamic behaviors [such as Hopf bifurcation (40) and chaos (41)] and various spatiotemporal patterns [such as irregular waves, target waves, traveling waves, and spiral waves (42, 43)] will appear when the system is subjected to different kinds and degrees of external stimulus. In the neural system, these spatiotemporal patterns are closely related to brain learning, memory, and information transmission. When the brain stores memory, the continuous firing rate of individual neurons shows a hierarchical change and the neurons show a strong temporal dynamic pattern and heterogeneity (33). Many factors contribute to the formation of short-term memory. Short-term memory does not form when the external stimulus is not sufficiently large (Figure 3). It is worth noting that neuronal activity is not only affected by external stimuli but also closely related to the interaction between nodes. The pattern remains flat when the external stimulus is large enough and the correlation degree of neurons is small. That is, short-term memory will not form (Figures 4A, B, 5A, B). When p increased to 0.01, neurons in the memory function areas fired, and the brain formed more vague memories (Figures 4C, D, 5C, D). Zheng et al. (37) found that neurons exhibit different pattern dynamics with the change of network connection probability p in the study of the HR model. This conclusion is also confirmed in the study of multilayer networks. Under the same degree of stimulation, if the number of neurons with the same functional attributes is different, the state of the neural network varies greatly (Figures 4B–D).

The physiological environment in which neurons work is always full of noise. From the above analysis, we can see that when $D_v = 9, p = 0.001$ is taken, the neurons are always in resting state (Figure 5A). To investigate the robustness of noise to the current results, we add Gaussian white noise to the multilayer FHN network model. The noise intensity np about u is used as the control

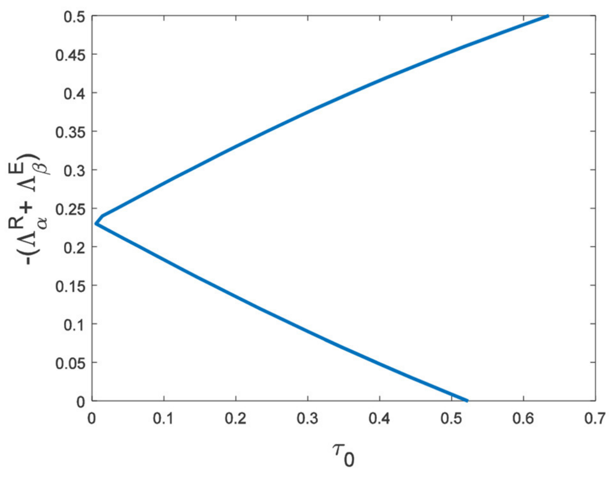


FIGURE 8
The relationship between τ_0 and eigenvalues of $R\square E$.

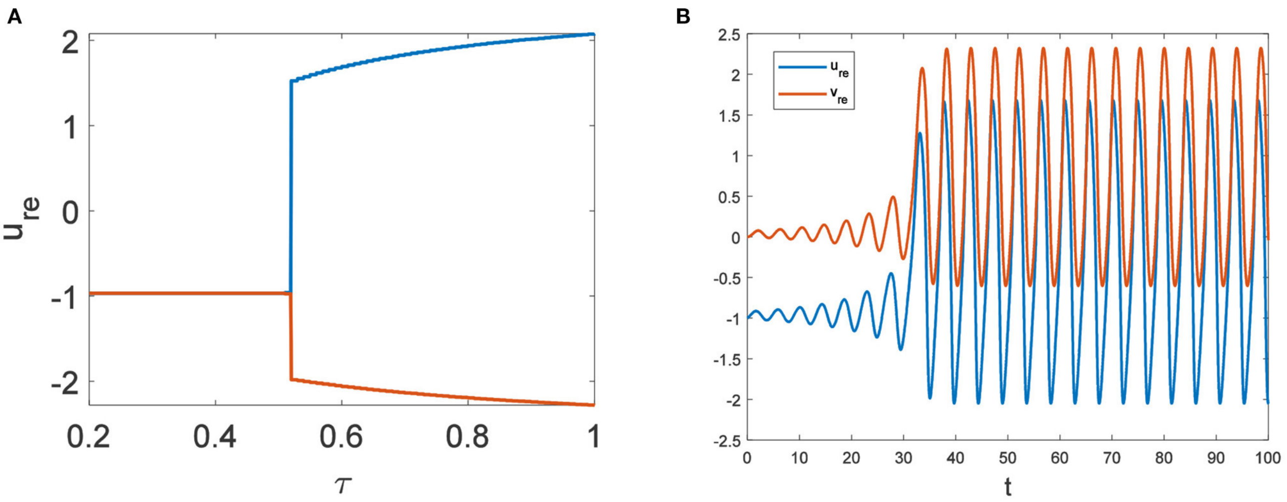


FIGURE 7
$D_u = D_v = 0$. **(A)** Bifurcation diagram about τ. The bifurcation point is $\tau = 0.5227$. **(B)** The time series diagram with $\tau = 0.6$.

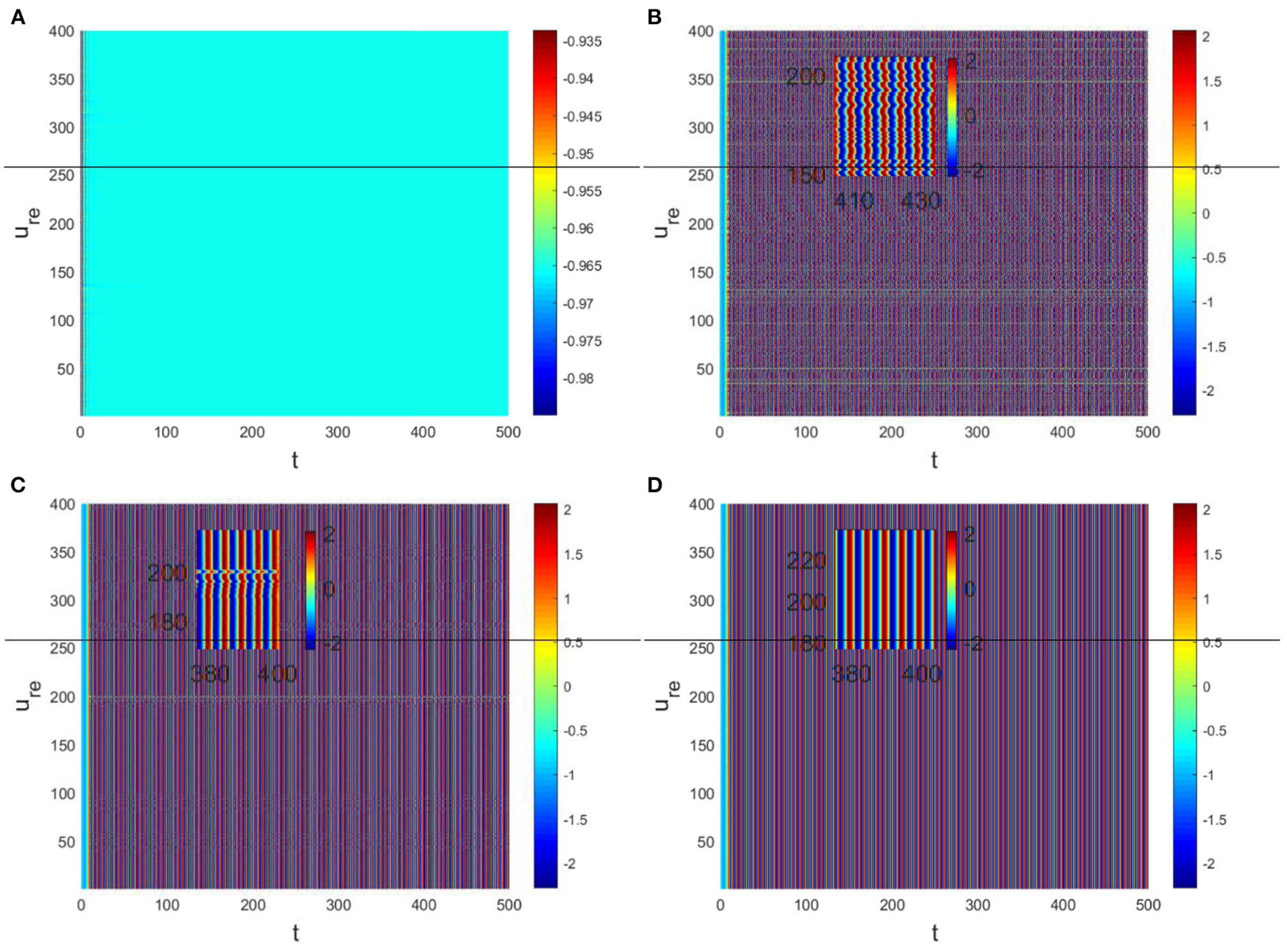


FIGURE 9
Pattern with $D_u = 0.01,\ D_v = 8$. **(A)** $\tau = 0.3,\ p = 0.01$. **(B)** $\tau = 1,\ p = 0.01$. **(C)** $\tau = 1,\ p = 0.1$. **(D)** $\tau = 1,\ p = 0.3$.

parameter. We find that the system is robust when $np < O(10^{-5})$; when $np > O(10^{-5})$, the neurons are excited and the short-term memory is vague (Figure 6).

In the neural system, synapses can regulate the release of excitatory neurotransmitters of membrane potential or mediators through delayed feedback, so the response and transmission of signals will be delayed. Time delay affects the generation of bifurcation and phase synchronization between neurons, which affects the brain's memory function (27). Next, we explore the effect of time delay on neuronal activity. The transition of neurons from resting state to firing state is always accompanied by bifurcation behavior. Action potential exceeds the threshold when the time delay is greater than $\tau_0 = 0.5227$, regardless of the influence of the network (Figure 7). $c_1(0) = 0.021 - 0.2729\imath,\ \mu_2 = -0.021,\ \beta_2 = 0.042,\ T_2 = 0.162$ can be found in Equation (27). Namely, the system generates subcritical Hopf bifurcation (similarly, we can get the supercritical Hopf bifurcation). From Figure 8, the network will affect the value of τ_0. In the study of the delayed neural network model, Zhao et al. (44) also found that the regulation of delay time can effectively control the formation of the pattern. Under the fixed network topology, the transmembrane current changes the membrane potential of neurons to different degrees with the increase of time delay. To more intuitively observe the collective behavior of neurons, we sorted 400 neuron nodes. When the delay time reaches 1, multiple neurons fire synchronously and participate in memory simultaneously (Figures 9A, B). It is found that the larger p is, the more obvious the synchronization phenomenon is (Figures 9B–D). Namely, short-term memory is relatively clear.

4. Conclusion

The brain is the most important organ in the human body, and its structure is very complex, so we have to simplify it when modeling. In this paper, we use the FHN model, which is simple but can describe the neuronal activity to explain the principle of short-term memory generation. The brain is a functional network that requires multiple neurons to work together for short-term memory. The brain regions responsible for specific tasks change their activity when the brain is storing memory (45). And pattern formation and selection can effectively detect collective behavior in excitable neural networks (27). Firstly, we establish the FHN model on the Cartesian product network and analyze the conditions of Turing instability. In the simulation, we found that short-term memory does not form when the probability of external stimulation and network connection is small. We test the robustness of the current results with Gaussian white noise and find that the system is robust when $np < O(10^{-5})$. Short-term memory is formed when external stimuli, network connection probability, and noise reach a certain range. Because the pattern is not regular at this time, the short-term memory is blurred. Then we study the effect of time delay on short-term memory formation and find that short-term memory is formed when the delay time exceeds τ_0. Of course, neuronal activity is not only related to external stimuli but the topology of the network itself. When p and delay time reach a certain degree, the cluster dynamic behavior appear, and the pattern shows periodic phenomenon. At this time, the brain forms a relatively clear short-term memory. These results provide a new way to explain the principle of memory formation.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

JS made contributions to the conception or design of mathematical model, supervision, and funding acquistition. JW made contributions to writing, editing–original draft, and formal analysis. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by National Natural Science Foundation of China (12272135) and Basic Research Project of Universities in Henan Province (21zx009).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Wang Q, Lu Q, Chen G, Duan L, et al. Bifurcation and synchronization of synaptically coupled FHN models with time delay. *Chaos Solitons Fractals.* (2009) 39:918–25. doi: 10.1016/j.chaos.2007.01.061

2. Tehrani NF, Razvan M. Bifurcation structure of two coupled FHN neurons with delay. *Math Biosci.* (2015) 270:41–56. doi: 10.1016/j.mbs.2015.09.008

3. Zhen B, Xu J. Fold-Hopf bifurcation analysis for a coupled FitzHugh-Nagumo neural system with time delay. *Int J Bifurcat Chaos.* (2010) 20:3919–34. doi: 10.1142/S0218127410028112

4. Lin Y. Periodic oscillation analysis for a coupled FHN network model with delays. In: *Abstract and Applied Analysis. Vol. 2013.* Hindawi (2013).

5. Han F, Zhen B, Du Y, Zheng Y, Wiercigroch M. Global Hopf bifurcation analysis of a six-dimensional FitzHugh-Nagumo neural network with delay by a synchronized scheme. *Discrete Continuous Dyn Syst B.* (2011) 16:457. doi: 10.3934/dcdsb.2011.16.457

6. Yu D, Wang G, Ding Q, Li T, Jia Y. Effects of bounded noise and time delay on signal transmission in excitable neural networks. *Chaos Solitons Fractals.* (2022) 157:111929. doi: 10.1016/j.chaos.2022.111929

7. Gan CB, Matjaz P, Qing-Yun W. Delay-aided stochastic multiresonances on scale-free FitzHugh-Nagumo neuronal networks. *Chin Phys B.* (2010) 19:040508. doi: 10.1088/1674-1056/19/4/040508

8. Zeng C, Zeng C, Gong A, Nie L. Effect of time delay in FitzHugh-Nagumo neural model with correlations between multiplicative and additive noises. *Physica A.* (2010) 389:5117–27. doi: 10.1016/j.physa.2010.07.031

9. Bashkirtseva I, Ryashko L. Analysis of excitability for the FitzHugh-Nagumo model via a stochastic sensitivity function technique. *Phys Rev E.* (2011) 83:061109. doi: 10.1103/PhysRevE.83.061109

10. Rajagopal K, Jafari S, Moroz I, Karthikeyan A, Srinivasan A. Noise induced suppression of spiral waves in a hybrid FitzHugh-Nagumo neuron with discontinuous resetting. *Chaos.* (2021) 31:073117. doi: 10.1063/5.0059175

11. Iqbal M, Rehan M, Hong KS. Robust adaptive synchronization of ring configured uncertain chaotic FitzHugh-Nagumo neurons under direction-dependent coupling. *Front Neurorobot.* (2018) 12:6. doi: 10.3389/fnbot.2018.00006

12. Feng P, Wu Y, Zhang J. A route to chaotic behavior of single neuron exposed to external electromagnetic radiation. *Front Comput Neurosci.* (2017) 11:94. doi: 10.3389/fncom.2017.00094

13. Liu H, Ge B. Turing instability of periodic solutions for the Gierer-Meinhardt model with cross-diffusion. *Chaos Solitons Fractals.* (2022) 155:111752. doi: 10.1016/j.chaos.2021.111752

14. Ghorai S, Poria S. Turing patterns induced by cross-diffusion in a predator-prey system in presence of habitat complexity. *Chaos Solitons Fractals.* (2016) 91:421–9. doi: 10.1016/j.chaos.2016.07.003

15. Lin J, Xu R, Li L. Turing-Hopf bifurcation of reaction-diffusion neural networks with leakage delay. *Commun Nonlinear Sci Num Simulat.* (2020) 85:105241. doi: 10.1016/j.cnsns.2020.105241

16. Mondal A, Upadhyay RK, Mondal A, Sharma SK. Emergence of Turing patterns and dynamic visualization in excitable neuron model. *Appl Math Comput.* (2022) 423:127010. doi: 10.1016/j.amc.2022.127010

17. Qu M, Zhang C. Turing instability and patterns of the FitzHugh-Nagumo model in square domain. *J Appl Anal Comput.* (2021) 11:1371–390. doi: 10.11948/20200182

18. Zheng Q, Shen J. Turing instability induced by random network in FitzHugh-Nagumo model. *Appl Math Computat.* (2020) 381:125304. doi: 10.1016/j.amc.2020.125304

19. Carletti T, Nakao H. Turing patterns in a network-reduced FitzHugh-Nagumo model. *Phys Rev E.* (2020) 101:022203. doi: 10.1103/PhysRevE.101.022203

20. Lei L, Yang J. Patterns in coupled FitzHugh-Nagumo model on duplex networks. *Chaos Solitons Fractals.* (2021) 144:110692. doi: 10.1016/j.chaos.2021.110692

21. Hu J, Zhu L. Turing pattern analysis of a reaction-diffusion rumor propagation system with time delay in both network and non-network environments. *Chaos Solitons Fractals.* (2021) 153:111542. doi: 10.1016/j.chaos.2021.111542

22. Yang W, Zheng Q, Shen J, Hu Q, Voit EO. Pattern dynamics in a predator-prey model with diffusion network. *Complex.* (2022) 2022:9055480. doi: 10.1155/2022/9055480

23. Ren Y, Sarkar A, Veltri P, Ay A, Dobra A, Kahveci T. Pattern discovery in multilayer networks. *IEEE/ACM Trans Comput Biol Bioinform*. (2021) 19:741–52. doi: 10.1109/TCBB.2021.3105001

24. Asllani M, Busiello DM, Carletti T, Fanelli D, Planchon G. Turing instabilities on Cartesian product networks. *Sci Rep*. (2015) 5:1–10. doi: 10.1038/srep12927

25. Tian C, Ruan S. Pattern formation and synchronism in an Allelopathic plankton model with delay in a network. *SIAM J Appl Dyn Syst*. (2019) 18:531–57. doi: 10.1137/18M1204966

26. Chen Z, Zhao D, Ruan J. Delay induced Hopf bifurcation of small-world networks. *Chin Ann Math B*. (2007) 28:453–62. doi: 10.1007/s11401-005-0300-z

27. Tang J. A review for dynamics of collective behaviors of network of neurons. *Sci China Technol Sci*. (2015) 58:2038–45. doi: 10.1007/s11431-015-5961-6

28. Bullmore E, Sporns O. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nat Rev Neurosci*. (2009) 10:186–98. doi: 10.1038/nrn2575

29. Wang XJ. Synaptic reverberation underlying mnemonic persistent activity. *Trends Neurosci*. (2001) 24:455–63. doi: 10.1016/S0166-2236(00)01868-3

30. Orhan AE, Ma WJ. A diverse range of factors affect the nature of neural representations underlying short-term memory. *Nat Neurosci*. (2019) 22:275–83. doi: 10.1038/s41593-018-0314-y

31. Murray JD, Bernacchia A, Roy NA, Constantinidis C, Romo R, Wang XJ. Stable population coding for working memory coexists with heterogeneous neural dynamics in prefrontal cortex. *Proc Natl Acad Sci USA*. (2017) 114:394–9. doi: 10.1073/pnas.1619449114

32. Bouchacourt F, Buschman TJ. A flexible model of working memory. *Neuron*. (2019) 103:147–60. doi: 10.1016/j.neuron.2019.04.020

33. Goldman MS. Memory without feedback in a neural network. *Neuron*. (2009) 61:621–34. doi: 10.1016/j.neuron.2008.12.012

34. Yao Y, Ma J. Weak periodic signal detection by sine-Wiener-noise-induced resonance in the FitzHugh-Nagumo neuron. *Cogn Neurodyn*. (2018) 12:343–9. doi: 10.1007/s11571-018-9475-3

35. Yu D, Zhou X, Wang G, Ding Q, Li T, Jia Y. Effects of chaotic activity and time delay on signal transmission in FitzHugh-Nagumo neuronal system. *Cogn Neurodyn*. (2022) 16:887–97. doi: 10.1007/s11571-021-09743-5

36. Zheng Q, Shen J, Xu Y. Spontaneous activity induced by gaussian noise in the network-organized fitzhugh-nagumo model. *Neural Plasticity*. (2020) 2020:6651441. doi: 10.1155/2020/6651441

37. Zheng Q, Shen J, Zhang R, Guan L, Xu Y. Spatiotemporal patterns in a general networked hindmarsh-rose model. *Front Physiol*. (2022) 13:936982. doi: 10.3389/fphys.2022.936982

38. Wang Z, Shi X. Electric activities of time-delay memristive neuron disturbed by Gaussian white noise. *Cogn Neurodyn*. (2020) 14:115–124. doi: 10.1007/s11571-019-09549-6

39. Van Gorder RA. Turing and Benjamin-Feir instability mechanisms in non-autonomous systems. *Proc R SocA*. (2020) 476:20200003. doi: 10.1098/rspa.2020.0003

40. Wouapi MK, Fotsin BH, Ngouonkadi EBM, Kemwoue FF, Njitacke ZT. Complex bifurcation analysis and synchronization optimal control for Hindmarsh-Rose neuron model under magnetic flow effect. *Cogn Neurodyn*. (2021) 15:315–47. doi: 10.1007/s11571-020-09606-5

41. Yang W. Bifurcation and dynamics in double-delayed Chua circuits with periodic perturbation. *Chin Phys B*. (2022) 31:020201. doi: 10.1088/1674-1056/ac1e0b

42. Rajagopal K, Ramadoss J, He S, Duraisamy P, Karthikeyan A. Obstacle induced spiral waves in a multilayered Huber-Braun (HB) neuron model. *Cogn Neurodyn*. (2022) 2022:1–15. doi: 10.1007/s11571-022-09785-3

43. Kang Y, Chen Y, Fu Y, Wang Z, Chen G. Formation of spiral wave in Hodgkin-Huxley neuron networks with Gamma-distributed synaptic input. *Commun Nonlinear Sci Num Simulat*. (2020) 83:105112. doi: 10.1016/j.cnsns.2019.105112

44. Zhao H, Huang X, Zhang X. Turing instability and pattern formation of neural networks with reaction-diffusion terms. *Nonlinear Dyn*. (2014) 76:115–124. doi: 10.1007/s11071-013-1114-2

45. Baldassarre A, Lewis CM, Committeri G, Snyder AZ, Romani GL, Corbetta M. Individual variability in functional connectivity predicts performance of a perceptual task. *Proc Natl Acad Sci USA*. (2012) 109:3516–21. doi: 10.1073/pnas.1113148109

frontiers | Frontiers in Psychiatry

TYPE Original Research
PUBLISHED 28 April 2023
DOI 10.3389/fpsyt.2023.1131769

OPEN ACCESS

EDITED BY
Jianzhong Su,
University of Texas at Arlington,
United States

REVIEWED BY
Bin Wu,
Beijing University of Posts and Telecommunications (BUPT),
China
Xiaojie Chen,
University of Electronic Science and Technology of China,
China

*CORRESPONDENCE
Wen-Xu Wang
✉ wenxuwang@bnu.edu.cn
Boyu Zhang
✉ zhangby@bnu.edu.cn

SPECIALTY SECTION
This article was submitted to Computational Psychiatry, a section of the journal Frontiers in Psychiatrys

RECEIVED 26 December 2022
ACCEPTED 10 March 2023
PUBLISHED 28 April 2023

CITATION
Li J-Y, Wu W-H, Li Z-Z, Wang W-X and Zhang B (2023) Data-driven evolutionary game models for the spread of fairness and cooperation in heterogeneous networks. *Front. Psychiatry* 14:1131769.
doi: 10.3389/fpsyt.2023.1131769

Data-driven evolutionary game models for the spread of fairness and cooperation in heterogeneous networks

Jing-Yi Li[1,2,3], Wen-Hao Wu[1], Ze-Zheng Li[1], Wen-Xu Wang[1,4]* and Boyu Zhang[5]*

[1]School of Systems Science, Beijing Normal University, Beijing, China, [2]CSSC Intelligent Innovation Research Institute, Beijing, China, [3]CSSC System Engineering Research Institute, Beijing, China, [4]Chinese Institute for Brain Research, Beijing, China, [5]Laboratory of Mathematics and Complex Systems, Ministry of Education, School of Mathematical Sciences, Beijing Normal University, Beijing, China

Unique large-scale cooperation and fairness norms are essential to human society, but the emergence of prosocial behaviors is elusive. The fact that heterogeneous social networks prevail raised a hypothesis that heterogeneous networks facilitate fairness and cooperation. However, the hypothesis has not been validated experimentally, and little is known about the evolutionary psychological basis of cooperation and fairness in human networks. Fortunately, research about oxytocin, a neuropeptide, may provide novel ideas for confirming the hypothesis. Recent oxytocin-modulated network game experiments observed that intranasal administration of oxytocin to a few central individuals significantly increases global fairness and cooperation. Here, based on the experimental phenomena and data, we show a joint effect of social preference and network heterogeneity on promoting prosocial behaviors by building evolutionary game models. In the network ultimatum game and the prisoner's dilemma game with punishment, inequality aversion can lead to the spread of costly punishment for selfish and unfair behaviors. This effect is initiated by oxytocin, then amplified *via* influential nodes, and finally promotes global cooperation and fairness. In contrast, in the network trust game, oxytocin increases trust and altruism, but these effects are confined locally. These results uncover general oxytocin-initiated mechanisms underpinning fairness and cooperation in human networks.

KEYWORDS

cooperation, fairness, inequality aversion, evolutionary game theory, social network

1. Introduction

Humans are self-organized to form a variety of social networks, upon which large-scale cooperative behaviors among genetically unrelated individuals persist (1–3). Human cooperation is crucial to the success of the human species and discriminates them from other species (4–6). To maintain cooperation, a preference for fairness in resource sharing is imperative and becomes a social norm (7–11). Despite significant progress in understanding the incentives of cooperation and fairness in spite of the temptation to be selfish, such as reciprocity and reputation (12–15), the emergence and evolution of cooperation and fairness in structured populations remain puzzling (16, 17).

Many efforts have attempted to interpret the effect of social networks on cooperation and fairness, among which a promising hypothesis is enlightened by a discovery in the field of complex networks (18–22). Much empirical evidence demonstrates that a large number of social and economic networks are heterogeneous, consisting of a small fraction of densely connected central nodes and a majority of sparsely connected peripheral nodes (23–26). It is believed that network heterogeneity plays a key role in cooperation and fairness, and several microscopic mechanisms based on social learning and natural selection have been proposed to explain the network reciprocity on cooperation (27–31). However, previous behavior experiments attempting to verify the network reciprocity hypothesis show negative results, and the influence of heterogeneous networks on prosocial behaviors becomes a debate (32, 33). How cooperation and fairness norms are enforced by social networks remains an outstanding problem.

Recent studies on social and behavioral neuroscience have explored the relationship between human behavior and oxytocin, a hypothalamic neuropeptide that has been associated with trust, fairness expectations, and social value representation (34–37). In particular, a placebo-controlled pharmacological study combining oxytocin and heterogeneous networks has shown that intranasal administration of oxytocin to a few central individuals can enhance global cooperation and fairness, but cannot affect global trust (38). Consequently, we hypothesize that heterogeneous networks indeed play a role in cooperation and fairness, but the effect of network heterogeneity is not prominent unless it is in coordination with individual differences in social preferences. Specifically, the leading effect of those prosocial individuals can be amplified by occupying influential nodes and could further exert a global impact on the whole network. It has been shown that oxytocin as a biological basis of prosocial behaviors accounts for the individual difference in social preference (37, 39, 40). We further hypothesize that individual differences are mainly differences in inequality aversion modulated by oxytocin and analyze the general mechanism underpinning fairness and cooperation in network environments initiated by oxytocin through a data-driven approach. Based on data from three network game experiments about fairness, cooperation, and trust (38), we build three evolutionary game dynamic models and reveal a remarkable joint effect of enhanced inequality aversion and network heterogeneity on global prosocial behaviors.

In the rest of this article, we first introduce the three base games and their network extensions adopted in the behavioral experiments (38): ultimatum game (UG) (41–43), two-stage prisoner's dilemma game with punishment (tPDG) (44–46) (a costly punishment stage is introduced on the basis of the classic prisoner's dilemma), and trust game (TG) (47–49). We then analyze strategy evolution in the three games on heterogeneous networks. Considering that individuals are bounded rational, we introduce (disadvantage) inequality aversion to the models (7). Specifically, inequality aversion means that people resist inequitable outcomes, and they are willing to give up some material payoffs to move in the direction of more equitable outcomes. We then construct utility matrices that incorporate both the material payoff and the influence of inequality aversion and analyze the evolution processes by replicator dynamic equations (50, 51). Based on the stability analysis of the dynamical systems and the real data of the experiments (38), the inequality aversion parameters are fitted. Our results show that in the UG and tPDG experiments, oxytocin can increase individual inequality aversion, thereby enhancing altruistic punishment, and this effect can be amplified and spread to the entire network through the network structure. In contrast, the trust enhanced by oxytocin fails to diffuse through the network structure to promote the level of prosociality in the TG network. In summary, our study can effectively explain the phenomenon in the behavioral experiments (38) and confirm that the leading effect caused by inequality aversion can be amplified by occupying influential nodes and further improve the level of cooperation and fairness of the whole network.

2. The ultimatum game on heterogeneous networks

2.1. Ultimatum game

The ultimatum game (UG) is a benchmark for studying fairness as a bounded rational behavior (7). Following Han et al. (25), we adopt a minimum acceptance offer (MAO) variant of UG that is simpler for playing on a network, but the essence of UG is not affected. In the networked UG, nine proposers and nine responders are placed at two kinds of nodes in a bipartite network, as shown in Figure 1A. Proposers have an identical neighborhood size with four responder neighbors. In contrast, there are two categories of responders, (i) three central responders with six proposer neighbors and (ii) six peripheral responders with three proposer neighbors. Every proposer connects to two central and two peripheral responders to ensure an unbiased influence from both categories. In each round, a proposer makes a single offer, resource p for each of his/her responder neighbors; a responder claims a single minimum amount, resource q that she/he can accept for all neighbors. Proposers and responders make decisions simultaneously, and every pair of connected subjects shares a fixed amount of resources. For each pair, if $p \geq q$, they make a deal and the proposer gets $1-p$ and the responder acquires p. If $p < q$, both get nothing. To control the effect of profit inequality resulting from heterogeneous connections, the actual payoff of subjects in a round is the average over their pairs of games.

2.2. Utility matrix

We classify the behaviors of proposers to be two categories: rational (R) with self-interest in payoffs or fairness (F) with fair sharing (52). To simplify our analyses, we assume the polarization of the two categories, i.e., rational proposers offer a small number of resources, s (with $s < 0.5$), to responders, and fair proposers offer 50% of resources to responders. Akin to proposers, we classify responders as cooperation (C, acquire any proposals not less than s) and defect (D, reject any proposals). In combination with the influence of disadvantage inequality aversion, we can define the utility of subjects with respect to both payoffs and inequality aversion and construct a utility matrix. Specifically, we assume that the utility of responders resulting from inequality aversion is proportional to the payoff difference, and an internal parameter characterizes the diversity of subjects in responding to inequality (7). The network inhomogeneity accounts for two representative responders: responders occupying central nodes and those occupying peripheral nodes. Thus, we define

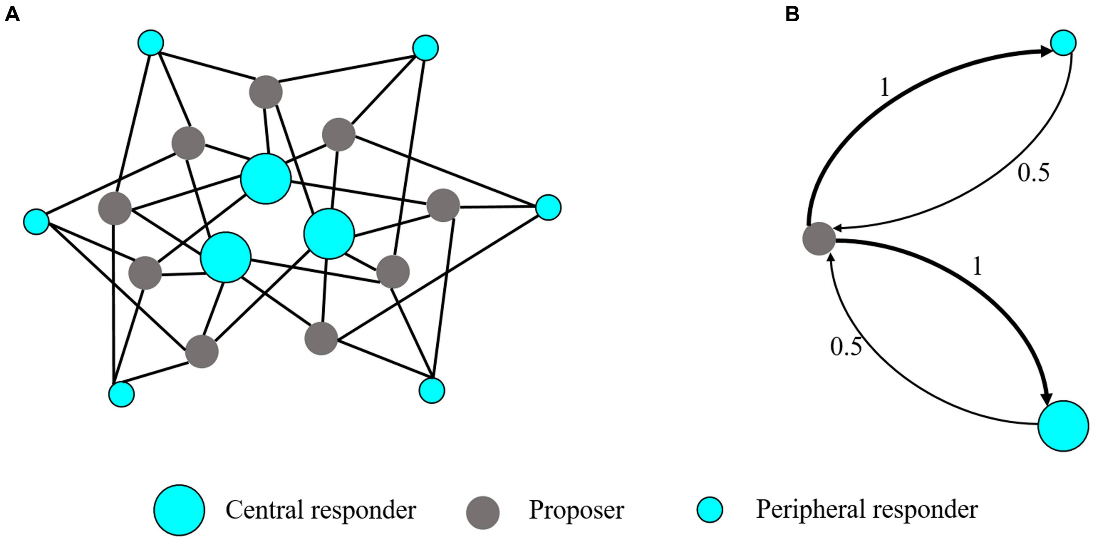


FIGURE 1
Network structure and mean-field approximation of the ultimatum game. **(A)** The network structure of UG. UG network has three central responders each connected to six proposer neighbors; six peripheral responders, each connected to three proposers and nine proposers each connected to four responders. The payoff of each subject is normalized by his/her number of neighbors. **(B)** The mean-field approximation of the UG network. The simplified system consists of three (types of) nodes, a proposer, a peripheral responder, and a central responder, representing three typical players in the network. The afferent arrows represent interaction with the neighbors of nodes. The thickness of the arrow and the value on the arrow represents the link weight.

TABLE 1 Utility matrix between proposer and responder.

(a) Proposer vs. central responder	C	D
R	$1-s, s-\alpha_1(1-2s)$	0,0
F	0.5,0.5	0,0
(b) Proposer vs. peripheral responder	C	D
R	$1-s, s-\alpha_2(1-2s)$	0,0
F	0.5,0.5	0,0

two utility matrices for the games between proposers and two types of responders (refer to Table 1), where $1-2s$ is the payoff difference between a rational proposer and a cooperative responder, α_1 and α_2 are the internal parameters for central and peripheral responders, respectively, which measure the degree of aversion to unfairness.

Our purpose is to estimate the values of internal parameters α_1 and α_2 that characterize the influence of oxytocin on the perception of inequality. To accomplish this goal, we employ replicator dynamics to model the evolution of subjects affected by their interactions. To enable analytical results, we reduce the network system with multi-type players based on the mean-field approximation method introduced in Zhang et al. (53) and Pei et al. (54). The basic idea of this method is to approximate the local network structure around a player (i.e., the distribution of different types of players in his/her neighborhood) with the global network structure (which can be derived from the frequencies of different types of edges) and approximate his/her local strategy distributions with the global strategy distributions. We note that this method can be applied to arbitrary networks, but in this article, we only focus on specific networks in Li et al. (38). As shown in Figure 1B, the simplified system consists of three nodes, a proposer, a peripheral responder, and a central responder, representing three typical players in the network. The links in the original network are converted to the interaction weights in the reduced network. The principle of the approximation is as follows:

- Because in the original network, each proposer connects to two central responders and two peripheral responders, in the reduced network the interaction weight from the central responder and the peripheral responder to the proposer is the same.
- Due to the fact that the payoff of each subject from playing with his/her neighbors is normalized by his/her number of neighbors, in the reduced network the sum of incoming link weights should be one.

Based on the approximation principle stemming from local interaction patterns, we can reasonably obtain the reduced network system in Figure 1B.

2.3. Replicator dynamics

To analyze the evolutionarily stable strategies for different types of nodes, we formulate replicator dynamics of the reduced network system. We denote the probability of proposers using the R strategy by ρ_R, the probability of central responders using the C strategy by ρ_C^C, and the probability of peripheral responders using the C strategy by ρ_C^P, respectively. In combination with the utility matrices, we can calculate the expected payoffs of subjects with different roles and strategies as follows:

$$\begin{cases} E(\mathrm{R}) = \frac{1}{2}\rho_{\mathrm{C}}^{\mathrm{C}}(1-s) + \frac{1}{2}\rho_{\mathrm{C}}^{\mathrm{P}}(1-s), \\ E(\mathrm{F}) = \frac{1}{4}\rho_{\mathrm{C}}^{\mathrm{C}} + \frac{1}{4}\rho_{\mathrm{C}}^{\mathrm{P}}, \\ E^{\mathrm{C}}(\mathrm{C}) = \rho_{\mathrm{R}}\left[s - \alpha_1(1-2s)\right] + \frac{1}{2}(1-\rho_{\mathrm{R}}), \\ E^{\mathrm{C}}(\mathrm{D}) = 0, \\ E^{\mathrm{P}}(\mathrm{C}) = \rho_{\mathrm{R}}\left[s - \alpha_2(1-2s)\right] + \frac{1}{2}(1-\rho_{\mathrm{R}}), \\ E^{\mathrm{P}}(\mathrm{D}) = 0, \end{cases} \tag{1}$$

where $E(\mathrm{R})$ and $E(\mathrm{F})$ are the expected payoffs of proposers with R and F strategies, respectively, $E^{\mathrm{C}}(\mathrm{C})$ and $E^{\mathrm{C}}(\mathrm{D})$ are the expected payoffs of central responders with C and D strategies, respectively, and $E^{\mathrm{P}}(\mathrm{C})$ and $E^{\mathrm{P}}(\mathrm{D})$ are the expected payoffs of peripheral responders with C and D strategies, respectively.

Thus, the replicator dynamics for the three types of nodes in the reduced network can be formulated as follows:

$$\begin{cases} \frac{d\rho_{\mathrm{R}}}{dt} = \rho_{\mathrm{R}}\left(E(\mathrm{R}) - \bar{E}\right) = \rho_{\mathrm{R}}(1-\rho_{\mathrm{R}})\left(E(\mathrm{R}) - E(\mathrm{F})\right), \\ \frac{d\rho_{\mathrm{C}}^{\mathrm{C}}}{dt} = \rho_{\mathrm{C}}^{\mathrm{C}}\left(E^{\mathrm{C}}(\mathrm{C}) - \overline{E^{\mathrm{C}}}\right) = \rho_{\mathrm{C}}^{\mathrm{C}}\left(1-\rho_{\mathrm{C}}^{\mathrm{C}}\right)\left(E^{\mathrm{C}}(\mathrm{C}) - E^{\mathrm{C}}(\mathrm{D})\right), \\ \frac{d\rho_{\mathrm{C}}^{\mathrm{P}}}{dt} = \rho_{\mathrm{C}}^{\mathrm{P}}\left(E^{\mathrm{P}}(\mathrm{C}) - \overline{E^{\mathrm{P}}}\right) = \rho_{\mathrm{C}}^{\mathrm{P}}\left(1-\rho_{\mathrm{C}}^{\mathrm{P}}\right)\left(E^{\mathrm{P}}(\mathrm{C}) - E^{\mathrm{P}}(\mathrm{D})\right), \end{cases} \tag{2}$$

where $\bar{E}$, $\overline{E^{\mathrm{C}}}$, and $\overline{E^{\mathrm{P}}}$ are the mean expected payoffs of proposers, central responders, and peripheral responders, respectively.

2.4. Stability analysis

The replicator dynamics do not have interior fixed points and have eight boundary fixed points $\left(\rho_{\mathrm{R}}, \rho_{\mathrm{C}}^{\mathrm{C}}, \rho_{\mathrm{C}}^{\mathrm{P}}\right)$, namely, (0, 0, 0), (0, 0, 1), (0, 1, 0), (0, 1, 1), (1, 0, 0), (1, 0, 1), (1, 1, 0), (1, 1, 1), where stable boundary fixed points correspond ESS of the game. We then implement stability analysis for each of the boundary fixed points (see SI for details). In general, Eq. (2) can have multiple stable fixed points. Since R is a dominant strategy for both types of proposers, we are more interested in the stable point with $\left(\rho_{\mathrm{C}}^{\mathrm{C}}, \rho_{\mathrm{C}}^{\mathrm{P}}\right) = (1,1)$. In this case, the only possible stable point is (1, 1, 1), where at this point proposers are rational and responders are cooperative.

Finally, we estimate the values of α_1, α_2, and s at (1, 1, 1) from the experimental data. From the stability condition, we have

$$\begin{cases} s < 0.5, \\ \frac{\alpha_1}{1+2\alpha_1} < s, \\ \frac{\alpha_2}{1+2\alpha_2} < s, \end{cases} \tag{3}$$

For convenience, let $q_1 = \frac{\alpha_1}{1+2\alpha_1}$ and $q_2 = \frac{\alpha_2}{1+2\alpha_2}$. Intuitively, q_1 (or q_2) represents the acceptance threshold of the central (or peripheral) responders adjacent to the proposer, where offers lower than the threshold will be rejected due to inequality aversion. Thus, we have

$$\begin{cases} \alpha_1 = \frac{q_1}{1-2q_1}, \\ \alpha_2 = \frac{q_2}{1-2q_2}. \end{cases} \tag{4}$$

Eq. (4) implies that the values of the inequality aversion parameters α_1 and α_2 can be estimated from the minimum acceptance offers q_1 and q_2. Here, we use data from the UG experiment in the study (38) to fit the parameters. The central nodes were given oxytocin or placebo in the experiment (the settings are the same in the following tPDG and TG experiments). The experimental group (administered oxytocin, OT) and the control group (administered placebo, PL) generated two sets of data, respectively. We use the mean minimum acceptance offers over 60 rounds of the central (or peripheral) responders adjacent to proposers to estimate q_1 (or q_2; see Table 2). The estimated values of α_1 and α_2 for OT and PL groups are shown in Table 3.

Not surprisingly, α_1 of the OT group is greater than those of the PL group, which implies that oxytocin indeed promotes inequality aversion. Interestingly, α_2 of the OT group is also higher. It indicates that oxytocin not only increases the inequality aversion of the central nodes but also spreads this influence to the entire network. Subsequently, we can predict the offer s of proposers determined by α_1 and α_2 based on our model. Note that to guarantee a deal with responders, a rational proposer's offer is confined by the condition $s = \max\{q_1, q_2\}$ (25). The predicted values of s for both OT and PL groups are shown in Figure 2.

The theoretical predictions are in good agreement with the experiment results, indicating that our model is effective. In particular, our model shows that the inequality aversion of the central responders and peripheral responders can be directly or indirectly increased by oxytocin. This is due to a subtle network effect, inequity aversion of central responders initiated by OT, self-interest of proposers induced by loss aversion, and conditional fairness of peripheral responders, which together constitute a mechanism underpinning the prosocial behaviors. Specifically, due to the endowment effect and loss aversion (55, 56), proposers use the best response strategy to maximize their payoffs and regard the offer to responders as a loss and often match the maximum q in their neighbors attempting to make all deals (25). The fact that responders refuse low offers because of inequity aversion resembles costly punishment to proposers. OT stimulates inequity aversion of central responders and imposes more punishment threats to unfair proposers. Despite the insignificant effect of OT exerting on only a small fraction of subjects, the local effect is amplified by central

TABLE 2 Estimated values of q_1 and q_2 in OT and PL groups.

	Central responder (q_1)	Peripheral responder(q_2)
OT	0.47	0.49
PL	0.43	0.47

TABLE 3 Estimated values of α_1 and α_2 in OT and PL groups.

α	Central responder (α_1)	Peripheral responder(α_2)
OT	8.73	19.61
PL	3.17	7.30

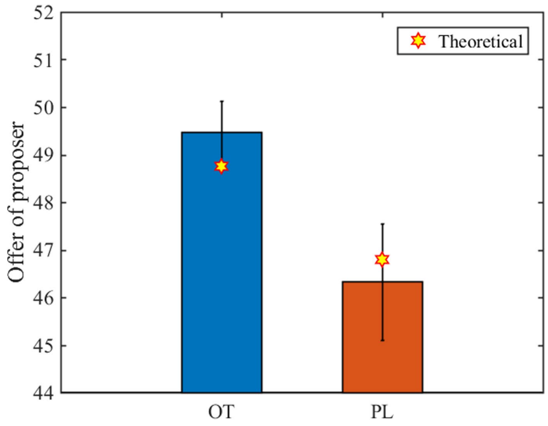


FIGURE 2
Experimental and predicted results of the proposers' offer. The blue bar and the red bar represent the mean value (all rounds of data) of the proposers' offer in the OT group and the PL group, respectively. Each error bar represents the standard deviation, and each yellow six-pointed star represents the theoretical value.

nodes with a larger number of connections. As a result, the central subjects become leaders in driving fairness behaviors *via* costly punishment.

Moreover, two complementary effects nudge network fairness. The first one is the complement among central nodes. Note that each proposer links to two central responders. Thus, only one of the central subjects who is qualified as a leader is sufficient to drive fairness of proposers who attempt to make all deals with their neighbors. The second complementary effect is ascribed to the conditional fairness of peripheral responders who increase their q insofar as their proposer neighbors increase their offers. In other words, the responders experience an inner conflict between advocating fairness and loss aversion, and the latter outweighs the former. The leaders help the responders overcome the obstacle of loss aversion and pursue fairness. Conditional fairness as compensation is important to sustain a high level of fairness during evolution in case of the fluctuation of the leader effect occasionally.

Taken together, OT initiates local costly sanctions on unfair behaviors by increasing the inequity aversion of subjects. The local effect is amplified by network heterogeneity and further assisted by conditional fairness and network complementary effects. Finally, the threat of punishment diffuses in the network and a high level of global fairness emerges. The mechanism underpinning network fairness enlightens us to explore network cooperation with costly punishment. We speculate that OT plays a similar role in costly punishment for selfish behaviors, and in combination with subtle network effects, cooperation could be fostered. We next analyze the behavioral evolution of the two-stage prisoner's dilemma game (tPDG) on the heterogeneous network by building the model to validate our hypothesis.

3. The prisoner's dilemma game with costly punishment on heterogeneous networks

3.1. Two-stage Prisoner's dilemma game

In the network of tPDG, there are two categories of nodes, three central nodes and nine peripheral nodes (Figure 3A), where each central and peripheral node has eight and four neighbors, respectively. To balance the influence of both categories, each peripheral node connects two central and two peripheral nodes, and each central node connects two central and six peripheral nodes. There are two stages in each round. In stage I, subjects choose either cooperate (C) or defect (D), and play with their neighbors simultaneously. The payoffs between each pair of neighboring subjects are calculated according to the payoff matrix (Table 4), where $\hat{T} > \hat{R} > \hat{P} > \hat{S}$. Similar to UG, the actual payoff of each subject is the average over all pairs of games in a round. In stage II, subjects can opt to costly punish their neighbors choosing D in stage I (57). The cost and punishment are normalized by the neighborhood size.

3.2. Utility matrix

To simplify our analyses and modeling processes, we merge the two steps and make an expanded payoff matrix associated with four strategies, i.e., cooperate and not punish (C+N), cooperate and punish (C+P), defect and not punish (D+N), and defect and punish (D+P) (46). The payoff matrix is shown in Table 5, where C is the cost of punishing neighbors with D strategy and F is the fine of punishment.

We speculate that few subjects will employ the D+P strategy. This strategy is not only strictly dominated by D+N but also cognitive dissonant in the sense that defectors punish other defectors. Thus, the payoff matrix can be reduced to three dimensions. Similar to the scenario in the UG [note that an alternative explanation for rejection in UG is that the responder punishes proposers by paying s such that the proposer loses 1-s, see (46)], we assume that the motivation of punishment is inequality aversion, where the willingness to punish defectors is positively related to $F-C$ (i.e., the efficiency of punishment). Meanwhile, cooperators who are defected may not choose to punish, especially central players tended to exhibit choosing to cooperate without punishing others' defection in oxytocin (vs. placebo) network (38). We speculate that the underlying reason is a kind of altruism (i.e., maximum group benefit) and may be affected by oxytocin. We further assume that this effect is positively related to $R-P$, (i.e., the benefit of mutual cooperation minus mutual defection). Regarding both inequality aversion and dilemma aversion, we have the utility matrix (Table 6) for central subjects, where $\beta_1\left(\hat{R}-\hat{P}\right)$ is the increase of utility by avoiding mutual defections, β_1 is an internal parameter capturing the individual difference in dilemma aversion, $\alpha_1\left(\hat{F}-\hat{C}\right)$ captures the willingness of punishment because of inequality aversion, and the internal parameter α_1 measures the degree of inequality aversion that could be affected by oxytocin. For peripheral subjects, we can write a similar utility matrix

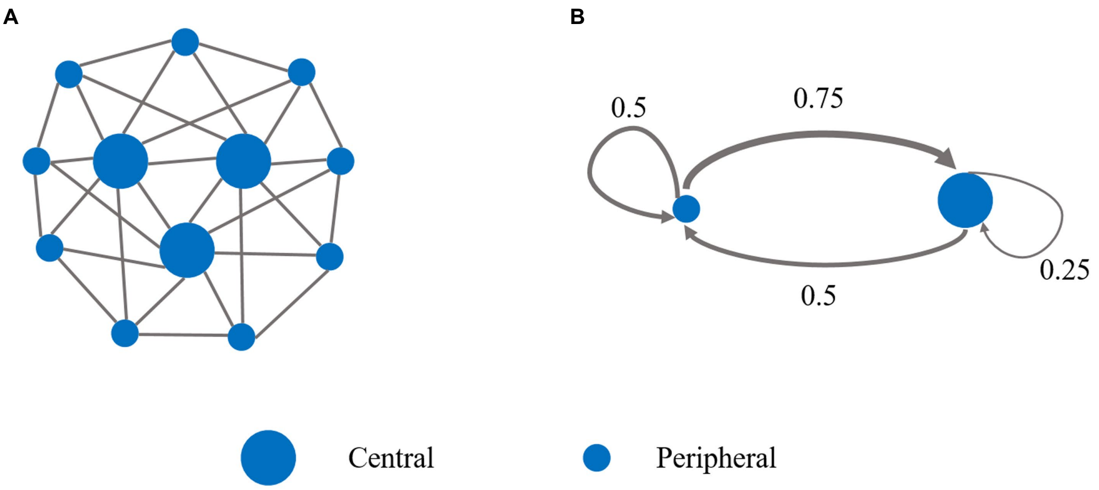


FIGURE 3
Network structure and mean-field approximation of the two-stage prisoner's dilemma game. **(A)** There are three central players and nine peripheral players in the tPDG network. Each central player was connected to eight neighbors (two central and six peripheral players), and each peripheral player was connected to only four neighbors (two central and two peripheral players). **(B)** The mean-field approximation of the tPDG network. The simplified system consists of (types of) two nodes, a peripheral node and a central node, representing two typical players in the network. The afferent arrows represent interaction with the neighbors of nodes. The thickness of the arrow and the value on the arrow represents the link weight.

TABLE 4 Payoff matrix in the prisoner's dilemma game.

	C	D
C	$\hat{R}, \hat{R}$	$\hat{S}, \hat{T}$
D	$\hat{T}, \hat{S}$	$\hat{P}, \hat{P}$

TABLE 5 Payoff matrix in the two-stage prisoner's dilemma game.

	C+N	C+P	D+N	D+P
C+N	$\hat{R}, \hat{R}$	$\hat{R}, \hat{R}$	$\hat{S}, \hat{T}$	$\hat{S}, \hat{T}$
C+P	$\hat{R}, \hat{R}$	$\hat{R}, \hat{R}$	$\hat{S}-\hat{C}, \hat{T}-\hat{F}$	$\hat{S}-\hat{C}, \hat{T}-\hat{F}$
D+N	$\hat{T}, \hat{S}$	$\hat{T}-\hat{F}, \hat{S}-\hat{C}$	$\hat{P}, \hat{P}$	$\hat{P}-\hat{F}, \hat{P}-\hat{C}$
D+P	$\hat{T}, \hat{S}$	$\hat{T}-\hat{F}, \hat{S}-\hat{C}$	$\hat{P}-\hat{C}, \hat{P}-\hat{F}$	$\hat{P}-\hat{C}-\hat{F}, \hat{P}-\hat{C}-\hat{F}$

(Table 6), where β_2 and α_2 represent the parameters of peripheral subjects.

Our aim is to estimate parameter values and reveal the effect of OT on the internal parameter α_1, α_2, β_1, and β_2. Analog to the case of network UG, we also use mean-field approximation to simplify our analyses. Because of the normalization of payoffs and punishment over neighbors of every subject, the original network can be reduced to a two-node graph with self-loops, as shown in Figure 3B.

In the original graph, a peripheral node connects to two other peripheral nodes and two central nodes, and a central node connects to six peripheral nodes and two other central nodes. Thus, the link weight of the self-loop of the peripheral node is 0.5, the same as the link weight from the central node to the peripheral node. The weight of the self-loop of the central node is $2/(2+6)=0.25$, and and the link weight from the peripheral node to the central node is 6/(2+6)=0.75.

3.3. Replicator dynamics

Next, we formulate replicator dynamics equations of the reduced network system. We denote the probabilities of central nodes using C+N, C+P, and D+N strategies by $\rho^{\mathrm{C}}_{\mathrm{C+N}}$, $\rho^{\mathrm{C}}_{\mathrm{C+P}}$, and $\rho^{\mathrm{C}}_{\mathrm{D+N}}$, respectively. Similarly, we denote the probabilities of peripheral nodes using C+N, C+P, and D+N strategies by $\rho^{\mathrm{P}}_{\mathrm{C+N}}$, $\rho^{\mathrm{P}}_{\mathrm{C+P}}$, and $\rho^{\mathrm{P}}_{\mathrm{D+N}}$, respectively. In combination with the utility matrices, we can calculate the expected payoffs of subjects with different roles and strategies, see SI for details. The replicator dynamics equations for the two nodes in the reduced network can be formulated as follows:

$$\begin{cases} \dfrac{d\rho^{\mathrm{C}}_{\mathrm{C+N}}}{dt} = \rho^{\mathrm{C}}_{\mathrm{C+N}}\left[E^{\mathrm{C}}(\mathrm{C+N}) - \overline{E^{\mathrm{C}}}\right], \\ \dfrac{d\rho^{\mathrm{C}}_{\mathrm{C+P}}}{dt} = \rho^{\mathrm{C}}_{\mathrm{C+P}}\left[E^{\mathrm{C}}(\mathrm{C+P}) - \overline{E^{\mathrm{C}}}\right], \\ \dfrac{d\rho^{\mathrm{P}}_{\mathrm{C+N}}}{dt} = \rho^{\mathrm{P}}_{\mathrm{C+N}}\left[E^{\mathrm{P}}(\mathrm{C+N}) - \overline{E^{\mathrm{P}}}\right], \\ \dfrac{d\rho^{\mathrm{P}}_{\mathrm{C+P}}}{dt} = \rho^{\mathrm{P}}_{\mathrm{C+P}}\left[E^{\mathrm{P}}(\mathrm{C+P}) - \overline{E^{\mathrm{P}}}\right]. \end{cases} \tag{5}$$

TABLE 6 Utility matrix in the two-stage prisoner's dilemma game.

(a) central player	C+N	C+P	D+N
C+N	$\hat{R}, \hat{R}$	$\hat{R}, \hat{R}$	$\hat{S}+\beta_1\left(\hat{R}-\hat{P}\right), \hat{T}$
C+P	$\hat{R}, \hat{R}$	$\hat{R}, \hat{R}$	$\hat{S}-\hat{C}+\alpha_1\left(\hat{F}-\hat{C}\right), \hat{T}-\hat{F}$
D+N	$\hat{T}, \hat{S}+\beta_1\left(\hat{R}-\hat{P}\right)$	$\hat{T}-\hat{F}, \hat{S}-\hat{C}+\alpha_1\left(\hat{F}-\hat{C}\right)$	$\hat{P}, \hat{P}$
(b) peripheral player	C+N	C+P	D+N
C+N	$\hat{R}, \hat{R}$	$\hat{R}, \hat{R}$	$\hat{S}+\beta_2\left(\hat{R}-\hat{P}\right), \hat{T}$
C+P	$\hat{R}, \hat{R}$	$\hat{R}, \hat{R}$	$\hat{S}-\hat{C}+\alpha_2\left(\hat{F}-\hat{C}\right), \hat{T}-\hat{F}$
D+N	$\hat{T}, \hat{S}+\beta_2\left(\hat{R}-\hat{P}\right)$	$\hat{T}-\hat{F}, \hat{S}-\hat{C}+\alpha_2\left(\hat{F}-\hat{C}\right)$	$\hat{P}, \hat{P}$

TABLE 7 Stable proportions of strategies in OT and PL groups.

	OT		PL	
	Central node	Peripheral node	Central node	Peripheral node
C+N	0.18	0.30	0.09	0.15
C+P	0.10	0.09	0.06	0.05
D+N	0.72	0.61	0.85	0.80

where $E^{C}(C+N)$ and $E^{C}(C+P)$ are the expected payoffs of the central node with C+N and C+P strategies, $E^{P}(C+N)$ and $E^{P}(C+P)$ are the expected payoffs of the peripheral node with C+N and C+P strategies, and $\overline{E^{C}}$ and $\overline{E^{P}}$ are the mean expected payoffs of the central node and peripheral node, respectively.

We then estimate the values of α_1, α_2, β_1, and β_2 by the virtue of experimental results. Note that the replicator dynamics are complicated with a large number of terms. This precludes us from deriving complete stability analyses for Eq. (5). Alternatively, we take the mean proportions of strategies of the last 20 rounds in experiments as the equilibrium points of the replicator dynamics, such that the parameter values in the dynamics can be estimated. Specifically, the (stable) proportions of strategies in OT and PL groups are shown in Table 7 (38). Thus, by inserting the equilibrium points into the replicator dynamics, we can solve the values of α_1, α_2, β_1, and β_2 for OT and PL groups, as shown in Table 8.

3.4. Stability analysis

Finally, we implement stability analysis to test if the equilibrium points in the experiments are indeed stable under Eq. (5). We formulate the Jacob matrix and calculated its eigenvalue using the estimated parameter values. We see that the real part corresponding to each eigenvalue of the Jacobi matrices is non-positive (see Supplementary Table S1), which indicates that the equilibrium state in the experiments is stable and can be achieved in our model. Thus, our evolutionary model is valid to model the evolution of cooperative behaviors in the prisoner's dilemma experiments with costly punishment (see Figure 4).

Our results indicate that oxytocin improves both the inequality aversion parameters α_1 and α_2 and the dilemma aversion parameters β_1, and β_2. Specifically, the costly punishment in stage II is analogous to rejecting unfair offers in UG, and OT triggers willingness to costly punishment by increasing inequity aversion of central subjects. The local punishment effect is amplified by central nodes and diffuses in the network by virtue of motivating conditional punishment of peripheral subjects. Actually, inspired by the sanction behaviors of central nodes, peripheral subjects' willingness to costly punishment in OT groups is significantly higher than that in PL groups.

4. The trust game on heterogeneous networks

4.1. Trust game

Due to the intensively studied effect of OT on trust, one may wonder whether trust (and altruism) increased by OT plays a role in fairness and cooperation in combination with inequity aversion and whether OT increases the trust of the whole network. In order to answer the questions, we analyze the trust game (TG) on the same heterogeneous network as that of UG (Figure 5A). Central and peripheral nodes are occupied by investors, and trustees have the same neighborhood size with four investor neighbors. Investors can choose

TABLE 8 Estimated values of α_1, α_2, β_1, and β_2 in OT and PL groups.

	OT	PL
Central node	$\alpha_1 = 3.13$ $\beta_1 = 1.13$	$\alpha_1 = 3.04$ $\beta_1 = 1.04$
Peripheral node	$\alpha_2 = 3.07$ $\beta_2 = 1.07$	$\alpha_2 = 3.02$ $\beta_2 = 1.02$

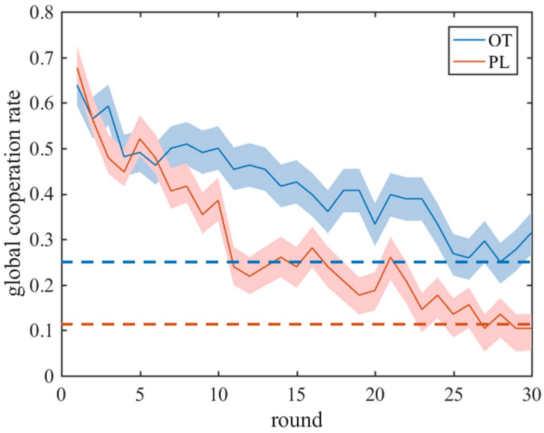


FIGURE 4
Experimental and predicted results of the global cooperation rate in tPDG. The solid lines represent the time evolution of the global cooperation rate in the iterated tPDG experiments and shaded areas represent standard error. The blue (red) lines represent OT (PL) group. The dotted lines represent the global cooperation rate at the stable fixed point of the replicator equations.

a certain proportion of the initial endowment as an investment. Trustees receive the investment increased by a certain multiple and decide how much to return to their investors. Therefore, both investors and trustees can obtain benefits through trust and reciprocity. However, trustees can exploit trust not to return any resources. In analogy with the settings in UG, the actual payoff of subjects in a round is the mean payoff over the number of their neighbors.

4.2. Utility matrix

In general, we classify the behaviors of investors into two categories: invest (I) in trustees from the initial endowment or do not invest (NI). Analogously, we classify trustees into two categories: return (R) a part of the investment to investors or do not return (NR). In addition, we take the altruistic preference of trustees into account to better model their behaviors and assume that the increase in the utility of trustees is proportional to the return. In the experiment, there are two categories of investors, those occupying central nodes and those occupying peripheral nodes. Thus, we define two utility matrices between a central investor and a trustee, and between a peripheral investor and a trustee, respectively.

The utility matrix of a central investor and a trustee is shown in Table 9, where T^{C} is the investment of a central investor with I strategy, r is the proportion of the investment that a trustee return to a central investor, g is the increase factor of investment ($g = 3$ in TG), rgT^{C} is the return from a trustee, $gT^{\mathrm{C}}(1-r)$ is the net gain of a trustee after returns $gT^{\mathrm{C}}r$, and the altruistic parameter λ ($\lambda > 0$) measures the willingness of return. In a similar manner, we define the utility matrix for a peripheral investor and a trustee (Table 9), where the superscript P denotes peripheral investors.

We aim to investigate the immediate effect of OT on investment T^{C} of central investors, and its possible indirect effect on investment T^{P} of peripheral investors, and the altruistic parameter λ of a trustee. We also use mean-field approximation to simplify our analyses. Because of the normalization of payoffs over neighbors of every subject, the original network can be reduced to a three-node graph, as shown in Figure 5B. The simplified system consists of a trustee, a peripheral investor, and a central investor. The links in the original network are converted to the interaction weights in the simplified graph, where the principle of the approximation is similar to the case of network UG. Based on the approximation principle stemming from local interaction patterns, we can obtain the simplified network system, as shown in Figure 5B.

4.3. Replicator dynamics

Next, we formulate replicator dynamics equations of the simplified network system. We denote the probability of trustees using the R strategy by ρ_{R}, the probability of central investors using the I strategy by $\rho_{\mathrm{I}}^{\mathrm{C}}$, and the probability of peripheral investors using the I strategy by $\rho_{\mathrm{I}}^{\mathrm{P}}$, respectively. According to the utility matrices, we can calculate the expected payoffs of subjects with different roles and strategies as follows:

$$\begin{cases} E(\mathrm{R}) = \frac{1}{2}\rho_{\mathrm{I}}^{\mathrm{C}}\left[gT^{\mathrm{C}}(1-r) + \lambda T^{\mathrm{C}} rg\right] + \frac{1}{2}\rho_{\mathrm{I}}^{\mathrm{P}}\left[gT^{\mathrm{P}}(1-r) + \lambda T^{\mathrm{P}} rg\right], \\ E(NR) = \frac{1}{2}\rho_{\mathrm{I}}^{\mathrm{C}} gT^{\mathrm{C}} + \frac{1}{2}\rho_{\mathrm{I}}^{\mathrm{P}} gT^{\mathrm{P}}, \\ E^{\mathrm{C}}(\mathrm{I}) = \rho_{\mathrm{R}}\left[\left(1-T^{\mathrm{C}}\right) + rgT^{\mathrm{C}}\right] + (1-\rho_{\mathrm{R}})\left(1-T^{\mathrm{C}}\right), \\ E^{\mathrm{C}}(NI) = 1, \\ E^{\mathrm{P}}(\mathrm{I}) = \rho_{\mathrm{R}}\left[\left(1-T^{\mathrm{P}}\right) + rgT^{\mathrm{P}}\right] + (1-\rho_{\mathrm{R}})\left(1-T^{\mathrm{P}}\right), \\ E^{\mathrm{P}}(NI) = 1. \end{cases} \tag{6}$$

where $E(\mathrm{R})$ and $E(\mathrm{NR})$ are the expected payoffs of trustees with R and NR strategies, $E^{\mathrm{C}}(\mathrm{I})$ and $E^{\mathrm{C}}(\mathrm{NI})$ are the expected payoffs of central investors with I and NI strategies, and $E^{\mathrm{P}}(\mathrm{I})$ and $E^{\mathrm{P}}(\mathrm{NI})$ are the expected payoffs of peripheral investors with I and NI strategies, respectively.

The three replicator dynamics equations for the three nodes in the simplified graph can be formulated as follows:

$$\begin{cases} \frac{d\rho_{\mathrm{R}}}{dt} = \rho_{\mathrm{R}}\left(E(\mathrm{R}) - \overline{E}\right) = \rho_{\mathrm{R}}(1-\rho_{\mathrm{R}})\left(E(\mathrm{R}) - E(NR)\right), \\ \frac{d\rho_I^{\mathrm{C}}}{dt} = \rho_{\mathrm{I}}^{\mathrm{C}}\left(E^{\mathrm{C}}(\mathrm{I}) - \overline{E^{\mathrm{C}}}\right) = \rho_{\mathrm{I}}^{\mathrm{C}}\left(1-\rho_{\mathrm{I}}^{\mathrm{C}}\right)\left(E^{\mathrm{C}}(\mathrm{I}) - E^{\mathrm{C}}(NI)\right), \\ \frac{d\rho_I^{\mathrm{P}}}{dt} = \rho_{\mathrm{I}}^{\mathrm{P}}\left(E^{\mathrm{P}}(\mathrm{I}) - \overline{E^{\mathrm{P}}}\right) = \rho_{\mathrm{I}}^{\mathrm{P}}\left(1-\rho_{\mathrm{I}}^{\mathrm{P}}\right)\left(E^{\mathrm{P}}(\mathrm{I}) - E^{\mathrm{P}}(NI)\right), \end{cases} \tag{7}$$

where $\overline{E}$, $\overline{E^{\mathrm{C}}}$, and $\overline{E^{\mathrm{P}}}$ are the mean expected payoffs of trustees, central investors, and peripheral investors, respectively.

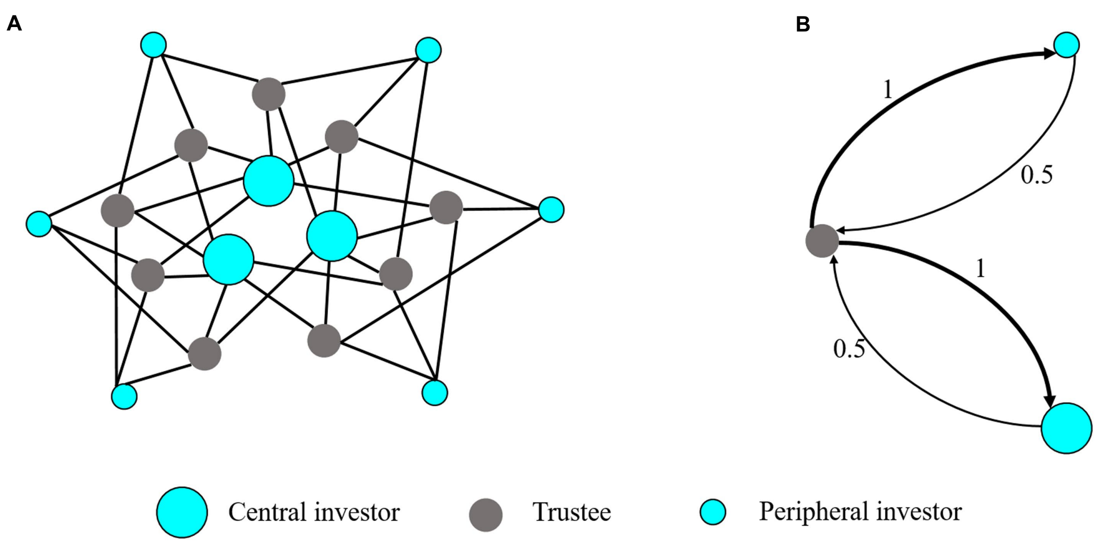


FIGURE 5
Network structure and mean-field approximation of the trust game. **(A)** The network structure of TG. The network has three central investors each connecting to six trustees; six peripheral investors, each connecting to three trustees, and nine trustees each connecting to four investors. **(B)** The mean-field approximation of the TG network. The simplified system consists of three (types of) nodes, a trustee, a peripheral investor, and a central investor, representing three typical players in the network. The afferent arrows represent interaction with the neighbors of nodes. The thickness of the arrow and the value on the arrow represents the connection link weight.

TABLE 9 Utility matrix between trustee and investor.

(a) Trustee vs. central investor	R	NR
I	$\left(1-T^{\mathrm{C}}\right)+rgT^{\mathrm{C}}, gT^{\mathrm{C}}(1-r)+\lambda T^{\mathrm{C}}rg$	$1-T^{\mathrm{C}}, gT^{\mathrm{C}}$
NI	1, 0	1, 0
(b) Trustee vs. peripheral investor	R	NR
I	$\left(1-T^{\mathrm{P}}\right)+rgT^{\mathrm{P}}, gT^{\mathrm{P}}(1-r)+\lambda T^{\mathrm{P}}rg$	$1-T^{\mathrm{P}}, gT^{\mathrm{P}}$
NI	1, 0	1, 0

4.4. Stability analysis

There exist nine possible equilibrium points $\left(\rho_{\mathrm{R}}, \rho_{\mathrm{I}}^{\mathrm{C}}, \rho_{\mathrm{I}}^{\mathrm{P}}\right)$ in the replicator dynamics equations, i.e., (0, 0, 0), (0, 0, 1), (0, 1, 0), (0, 1, 1), (1, 0, 0), (1, 0, 1), (1, 1, 0), (1, 1, 1), $\left(\frac{1}{rg}, 0, 0\right)$. We implement a stability analysis for each of the equilibrium points. For $\lambda < 1$, Eq. (7) has only one stable equilibrium, (0, 0, 0), but this equilibrium cannot explain all the experimental results. For $\lambda > 1$, Eq. (7) can have four possible stable states: (0, 0, 0), (1, 0, 1), (1, 1, 0), and (1, 1, 1). According to stability conditions, these stable states can be classified into three categories ($g = 3$) as follows:

(i) For $\mathrm{r} = 0$, (0, 0, 0) is stable;
(ii) For $\mathrm{r} = \frac{1}{g}$, (1, 0, 1) and (1, 1, 0) are stable;
(iii) For $\mathrm{r} > \frac{1}{g}$, (1, 1, 1) is stable.

Subsequently, we analyze the stable points of each group of experiments to classify these groups into three categories. There are nine groups in the OT experiments and 10 groups in the PL experiments. The stable point and the classification of each group can be found in Supplementary Tables S2, S3.

The classification and stable point of experimental results demonstrate that our model is valid to characterize the evolutionary features of the trust experiment. In particular, we can see that the behavior r of the trustee is not affected by the investment T of investors or the altruistic parameter λ, and only the relation between r and g influences the category of experimental behaviors. In other words, OT that directly affects T and λ values plays a negligible role in the behavior of trustees, such that the dynamics of the experiment is not affected by OT as well. It is worth noting that OT, indeed, enhances the investment T^{C} of central investors administrated OT by comparing with those of peripheral investors without inhaling OT (38). The results indicate that the effect of OT on enhancing the trust of investors is confined locally and cannot spread to other peripheral investors, due to the fact that the neighboring trustees of the central investors show no response to the generosity of the investors and shield the effect of OT. This finding is consistent with a pioneering experiment of one pair of investor and trustee, in which OT only affect the generosity of investors but is useless to trustees (58).

Taken together, locally administrated OT has no effect on the trust game experiments. This is mainly ascribed by the awarding from investors trigger by OT, which is not strong enough to motivate significant higher return of trustees. In contrast, in the UG and PD with costly punishment, the punishment stemming from inequality aversion triggered by locally administrated OT is effective to promote fairness and cooperation. In brief, group fairness and cooperative behavior can result from inequity aversion rather than trust.

5. Conclusion and discussion

Humans have a strong capacity to cooperate with genetically unrelated individuals. Yet because cooperation is exploitable by free-riding, when and how large-scale cooperation emerges and spreads through human social networks remains puzzling from both evolutionary and societal perspectives.

The effect of the heterogeneous network on group cooperation has always been a hot issue in related fields. Considering oxytocin is believed to be a neuropeptide with positive effects on prosocial behavior (e.g., positive effects on trust), the recent research conducted oxytocin-modulated network game experiments, including: the ultimatum game, the two-stage prisoner's dilemma game (with the costly punishment stage), and the trust game in heterogeneous networks, respectively, and found that the administration of oxytocin (vs. matching placebo) to central individuals can increase the level of cooperation and fairness in the whole network significantly. Here, in order to further explore the intrinsic mechanism of this experimental phenomenon, we analyzed the evolution process of three game experiments in heterogeneous networks by constructing evolutionary game dynamics, respectively.

Based on the experimental data, the parameter estimation on the analytical results of the evolutionary game models shows that oxytocin can significantly enhance the prosocial preferences of the central subjects in all three games. In the UG and tPDG models, the altruistic punishment caused by inequality aversion is amplified and diffused through the heterogeneous network structure, thereby promoting cooperation and fairness in the overall network.

However, no cascading effects of oxytocin-induced prosocial behavior were observed in repeated rounds of TG experiments that did not involve inequality aversion (38). Oxytocin can significantly increase the investment (trust level) of investors, which is equivalent to a reward for the trustee (the incentive effect of reward is far weaker than punishment). However, the investor may lack a mechanism for punishment, and the trustee is not threatened with punishment and thus will not increase his/her return. In our model, we find that the trustee's return ratio r is not affected by the investor's investment T , which can effectively explain the experimental results. Therefore, we can conclude that the rewarding effect of trust is not sufficient to generate prosocial utility and that the costly punishment caused by inequality aversion is more effective in promoting the level of fairness and cooperation in the social network. These results confirm our hypothesis and may also explain existing network-free findings on punishment and reward (46, 59).

Our study opens an avenue to uncover general oxytocin-initiated mechanisms underpinning fairness and cooperation in human society through building evolutionary game models. Our evolutionary game model is a network variant of the replicator dynamics. Replicator dynamics have been widely used to study the evolution of cooperation and fairness in social networks (17, 60–62). One implicit assumption of the replicator dynamics is that imitation only occurs among individuals of the same type. While in the game experiments (38), subjects were also informed of the choices and payoffs of other types of subjects. Thus, they may not make decisions based on local imitation. However, it is worth noting that the goal of our study is not to exactly reproduce individual behaviors in the game experiments, but rather to show that the results observed in the experiments can be achieved and are evolutionarily stable in simple evolutionary game models.

In addition, our study provides an effective means to quantitively estimate the effect of oxytocin on inequality aversion and trust based on the experimental data. Thus, a possible direction for future research is to design experiments with different quantities of oxytocin, and we believe that our method can contribute to measuring how the quantity of oxytocin affects different social preferences. In short, our study and its future extension provide a new perspective for understanding the relationship between neuropeptides and prosocial behaviors.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding authors.

Author contributions

J-YL performed the analysis. J-YL, W-XW, and BZ wrote the first draft of the manuscript. All authors contributed to conception and design of the study, manuscript revision, and read and approved the submitted version.

Funding

This study was supported by the National Natural Science Foundation of China (Nos. 11975049, 72131003, and 71922004), the Beijing Natural Science Foundation (No. Z220001), and the Collaborative Research Fund of the Chinese Institute for Brain Research, Beijing.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyt.2023.1131769/full#supplementary-material

References

1. Fehr E, Fischbacher U. The nature of human altruism. *Nature*. (2003) 425:785–91. doi: 10.1038/nature02043

2. Fehr E, Schurtenberger I. Normative foundations of human cooperation. *Nat Hum Behav*. (2018) 2:458–68. doi: 10.1038/s41562-018-0385-5

3. Jackson MO. *Social and Economic Networks*. Princeton, NJ: Princeton University Press (2010).

4. Hill KR, Walker RS, Božičević M, Eder J, Headland T, Hewlett B, et al. Co-residence patterns in hunter-gatherer societies show unique human social structure. *Science*. (2011) 331:1286–9. doi: 10.1126/science.1199071

5. Apicella CL, Silk JB. The evolution of human cooperation. *Curr Biol*. (2019) 29:R447–50. doi: 10.1016/j.cub.2019.03.036

6. Trivers RL. The evolution of reciprocal altruism. *Q Rev Biol*. (1971) 46:35–57. doi: 10.1086/406755

7. Fehr E, Schmidt KM. A theory of fairness, competition, and cooperation. *Q J Econ*. (1999) 114:817–68. doi: 10.1162/003355399556151

8. Bolton GE, Ockenfels A. ERC: a theory of equity, reciprocity, and competition. *Am Econ Rev*. (2000) 90:166–93. doi: 10.1257/aer.90.1.166

9. Dawes CT, Fowler JH, Johnson T, McElreath R, Smirnov O. Egalitarian motives in humans nature. *Nature*. (2007) 446:794–6. doi: 10.1038/nature05651

10. Gavrilets S, Richerson PJ. Collective action and the evolution of social norm internalization. *Proc Natl Acad Sci*. (2017) 114:6068–73. doi: 10.1073/pnas.1703857114

11. Fehr E, Fischbacher U. Social norms and human cooperation. *Trends Cogn Sci*. (2004) 8:185–90. doi: 10.1016/j.tics.2004.02.007

12. Fehr E, Fischbacher U, Gächter S. Strong reciprocity, human cooperation, and the enforcement of social norms. *Hum Nat*. (2002) 13:1–25. doi: 10.1007/s12110-002-1012-7

13. Falk A, Fischbacher U. A theory of reciprocity. *Games Econ Behav*. (2006) 54:293–315. doi: 10.1016/j.geb.2005.03.001

14. Rockenbach B, Milinski M. The efficient interaction of indirect reciprocity and costly punishment. *Nature*. (2006) 444:718–23. doi: 10.1038/nature05229

15. Lyle HF, Smith EA. The reputational and social network benefits of prosociality in an Andean community. *Proc Natl Acad Sci*. (2014) 111:4820–5. doi: 10.1073/pnas.1318372111

16. Sinha S, Ghosh S, Roy S. A pedestrian review of games on structured populations. *Int J Adv Eng Sci Appl Math*. (2019) 11:138–52. doi: 10.1007/s12572-018-0241-x

17. Szabó G, Fath G. Evolutionary games on graphs. *Phys Rep*. (2007) 446:97–216. doi: 10.1016/j.physrep.2007.04.004

18. Sinatra R, Iranzo J, Gomez-Gardenes J, Floria LM, Latora V, Moreno Y. The ultimatum game in complex networks. *J Stat Mech Theory Exp*. (2009) 2009:P09012. doi: 10.1088/1742-5468/2009/09/P09012

19. Iranzo J, Floria LM, Moreno Y, Sanchez A. Empathy emerges spontaneously in the ultimatum game: Small groups and networks. *PLoS One*. (2012) 7:e43781.

20. Oosterbeek H, Sloof R, Van De Kuilen G. Cultural differences in ultimatum game experiments: evidence from a meta-analysis. *Exp Econ*. (2004) 7:171–88. doi: 10.1023/B:EXEC.0000026978.14316.74

21. Rand DG, Nowak MA, Fowler JH, Christakis NA. Static network structure can stabilize human cooperation. *Proc Natl Acad Sci*. (2014) 111:17093–8. doi: 10.1073/pnas.1400406111

22. Henrich J, Ensminger J, McElreath R, Barr A, Barrett C, Bolyanatz A, et al. Markets, religion, community size, and the evolution of fairness and punishment. *Science*. (2010) 327:1480–4. doi: 10.1126/science.1182238

23. Barabási A-L, Albert R. Emergence of scaling in random networks. *Science*. (1999) 286:509–12. doi: 10.1126/science.286.5439.509

24. Pastor-Satorras R, Castellano C, Van Mieghem P, Vespignani A. Epidemic processes in complex networks. *Rev Mod Phys*. (2015) 87:925. doi: 10.1103/RevModPhys.87.925

25. Han X, Cao S, Shen Z, Zhang B, Wang W-X, Cressman R, et al. Emergence of communities and diversity in social networks. *Proc Natl Acad Sci*. (2017) 114:2887–91. doi: 10.1073/pnas.1608164114

26. Fortunato S, Bergstrom CT, Börner K, Evans JA, Helbing D, Milojević S, et al. Science of science. *Science*. (2018) 359:6379. doi: 10.1126/science.aao0185

27. Allen B, Lippner G, Chen Y-T, Fotouhi B, Momeni N, Yau S-T, et al. Evolutionary dynamics on any population structure. *Nature*. (2017) 544:227–30. doi: 10.1038/nature21723

28. Santos FC, Pacheco JM, Lenaerts T. Evolutionary dynamics of social dilemmas in structured heterogeneous populations. *Proc Natl Acad Sci*. (2006) 103:3490–4. doi: 10.1073/pnas.0508201103

29. Gächter S, Nosenzo D, Renner E, Sefton M. Who makes a good leader? Cooperativeness, optimism, and leading-by-example. *Econ Inq*. (2012) 50:953–67. doi: 10.1111/j.1465-7295.2010.00295.x

30. Apicella CL, Marlowe FW, Fowler JH, Christakis NA. Social networks and cooperation in hunter-gatherers. *Nature*. (2012) 481:497–501. doi: 10.1038/nature10736

31. Santos FC, Santos MD, Pacheco JM. Social diversity promotes the emergence of cooperation in public goods games. *Nature*. (2008) 454:213–6. doi: 10.1038/nature06940

32. Gracia-Lázaro C, Ferrer A, Ruiz G, Tarancón A, Cuesta JA, Sánchez A, et al. Heterogeneous networks do not promote cooperation when humans play a Prisoner's dilemma. *Proc Natl Acad Sci*. (2012) 109:12922–6. doi: 10.1073/pnas.1206681109

33. Maciejewski W, Fu F, Hauert C. Evolutionary game dynamics in populations with heterogenous structures. *PLoS Comput Biol*. (2014) 10:e1003567. doi: 10.1371/journal.pcbi.1003567

34. Carter CS. Oxytocin pathways and the evolution of human behavior. *Annu Rev Psychol*. (2014) 65:17–39. doi: 10.1146/annurev-psych-010213-115110

35. Declerck CH, Boone C, Pauwels L, Vogt B, Fehr E. A registered replication study on oxytocin and trust. *Nat Hum Behav*. (2020) 4:646–55. doi: 10.1038/s41562-020-0878-x

36. Liu Y, Li S, Lin W, Li W, Yan X, Wang X, et al. Oxytocin modulates social value representations in the amygdala. *Nat Neurosci*. (2019) 22:633–41. doi: 10.1038/s41593-019-0351-1

37. Stallen M, Rossi F, Heijne A, Smidts A, De Dreu CK, Sanfey AG. Neurobiological mechanisms of responding to injustice. *J Neurosci*. (2018) 38:2944–54. doi: 10.1523/JNEUROSCI.1242-17.2018

38. Li S, Ma S, Wang D, Zhang H, Li Y, Wang J, et al. Oxytocin and the punitive hub—dynamic spread of cooperation in human social networks. *J Neurosci*. (2022) 42:5930–43. doi: 10.1523/JNEUROSCI.2303-21.2022

39. Spengler FB, Scheele D, Marsh N, Kofferath C, Flach A, Schwarz S, et al. Oxytocin facilitates reciprocity in social communication. *Soc Cogn Affect Neurosci*. (2017) 12:1325–33. doi: 10.1093/scan/nsx061

40. Marsh N, Marsh AA, Lee MR, Hurlemann R. Oxytocin and the neurobiology of prosocial behavior. *Neuroscientist*. (2021) 27:604–19. doi: 10.1177/1073858420960111

41. Güth W, Schmittberger R, Schwarze B. An experimental analysis of ultimatum bargaining. *J Econ Behav Organ*. (1982) 3:367–88. doi: 10.1016/0167-2681(82)90011-7

42. Wang X, Chen X, Wang L. Evolutionary dynamics of fairness on graphs with migration. *J Theor Biol*. (2015) 380:103–14. doi: 10.1016/j.jtbi.2015.05.020

43. Zhang Y, Chen X, Liu A, Sun C. The effect of the stake size on the evolution of fairness. *Appl Math Comput*. (2018) 321:641–53. doi: 10.1016/j.amc.2017.11.013

44. Rapoport A, Chammah AM, Orwant CJ. *Prisoner's Dilemma: A Study in Conflict and Cooperation*. Ann Arbor, MI: University of Michigan Press (1965).

45. Dawes RM. Social dilemmas. *Annu Rev Psychol*. (1980) 31:169–93. doi: 10.1146/annurev.ps.31.020180.001125

46. Sigmund K, Hauert C, Nowak MA. Reward and punishment. *Proc Natl Acad Sci*. (2001) 98:10757–62. doi: 10.1073/pnas.161155698

47. Berg J, Dickhaut J, McCabe K. Trust, reciprocity, and social history. *Games Econ Behav*. (1995) 10:122–42. doi: 10.1006/game.1995.1027

48. Liu L, Chen X. Conditional investment strategy in evolutionary trust games with repeated group interactions. *Inf Sci*. (2022) 609:1694–705. doi: 10.1016/j.ins.2022.07.073

49. Sun K, Liu Y, Chen X, Szolnoki A. Evolution of trust in a hierarchical population with punishing investors. *Chaos, Solitons Fractals*. (2022) 162:112413. doi: 10.1016/j.chaos.2022.112413

50. Hofbauer J, Sigmund K. Evolutionary game dynamics. *Bull Am Math Soc*. (2003) 40:479–519. doi: 10.1090/S0273-0979-03-00988-1

51. Taylor PD, Jonker LB. Evolutionary stable strategies and game dynamics. *Math Biosci*. (1978) 40:145–56. doi: 10.1016/0025-5564(78)90077-9

52. Nowak MA, Page KM, Sigmund K. Fairness versus reason in the ultimatum game. *Science*. (2000) 289:1773–5. doi: 10.1126/science.289.5485.1773

53. Zhang B, Cao Z, Qin C-Z, Yang X. Fashion and homophily. *Oper Res*. (2018) 66:1486–97. doi: 10.1287/opre.2018.1744

54. Pei S, Cressman R, Zhang B. Dynamic games on networks with heterogeneous players. Available at SSRN: https://ssrn.com/abstract=4217352

55. Thaler R. Toward a positive theory of consumer choice. *J Econ Behav Organ*. (1980) 1:39–60. doi: 10.1016/0167-2681(80)90051-7

56. Tversky A, Kahneman D. Loss aversion in riskless choice: a reference-dependent model. *Q J Econ*. (1991) 106:1039–61. doi: 10.2307/2937956

57. Zhang BY, Li C, De Silva H, Bednarik P, Sigmund K. The evolution of sanctioning institutions: an experimental approach to the social contract. *Exp Econ*. (2014) 17:285–303. doi: 10.1007/s10683-013-9367-7

58. Kosfeld M, Heinrichs M, Zak PJ, Fischbacher U, Fehr E. Oxytocin increases trust in humans. *Nature*. (2005) 435:673–6. doi: 10.1038/nature03701

59. Wu J, Luan S, Raihani N. Reward, punishment, and prosocial behavior: recent developments and implications. *Curr Opin Psychol*. (2022) 44:117–23. doi: 10.1016/j.copsyc.2021.09.003

60. Ohtsuki H, Nowak MA. The replicator equation on graphs. *J Theor Biol*. (2006) 243:86–97. doi: 10.1016/j.jtbi.2006.06.004

61. Ohtsuki H, Nowak MA, Pacheco JM. Breaking the symmetry between interaction and replacement in evolutionary dynamics on graphs. *Phys Rev Lett*. (2007) 98:108106. doi: 10.1103/PhysRevLett.98.108106

62. Zhang B, Li C, Tao Y. Evolutionary stability and the evolution of cooperation on heterogeneous graphs. *Dyn Games Appl*. (2016) 6:567–79. doi: 10.1007/s13235-015-0146-2

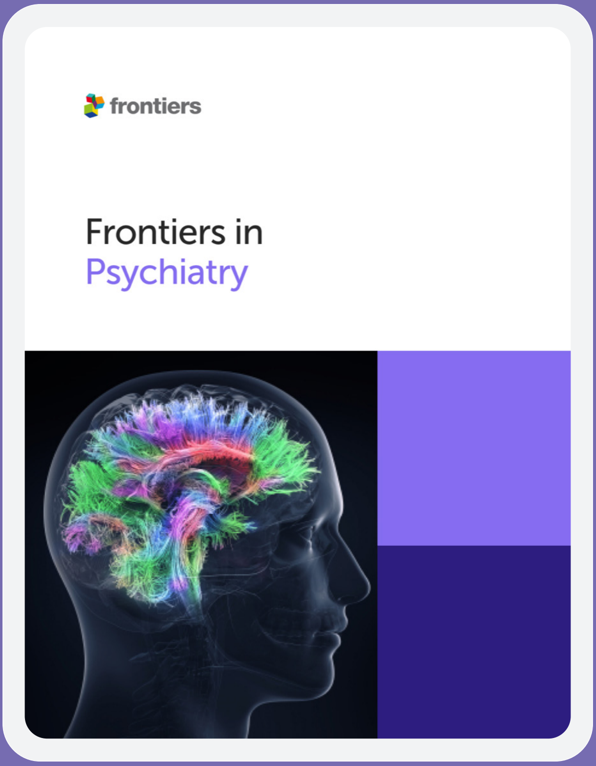
Frontiers in Psychiatry

Explores and communicates innovation in the field of psychiatry to improve patient outcomes

The third most-cited journal in its field, using translational approaches to improve therapeutic options for mental illness, communicate progress to clinicians and researchers, and consequently to improve patient treatment outcomes.

Discover the latest Research Topics

See more →



frontiers

Frontiers in Psychiatry

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact