ARTICLES
https://doi.org/10.1038/s41562-019-0804-2
nature human behaviour

Quantum reinforcement learning during human decision-making

Ji-An Li[1,2], Daoyi Dong[3], Zhengde Wei[1,4], Ying Liu[5], Yu Pan[6], Franco Nori[7,8] and Xiaochu Zhang[1,9,10,11]*

Classical reinforcement learning (CRL) has been widely applied in neuroscience and psychology; however, quantum reinforcement learning (QRL), which shows superior performance in computer simulations, has never been empirically tested on human decision-making. Moreover, all current successful quantum models for human cognition lack connections to neuroscience. Here we studied whether QRL can properly explain value-based decision-making. We compared 2 QRL and 12 CRL models by using behavioural and functional magnetic resonance imaging data from healthy and cigarette-smoking subjects performing the Iowa Gambling Task. In all groups, the QRL models performed well when compared with the best CRL models and further revealed the representation of quantum-like internal-state-related variables in the medial frontal gyrus in both healthy subjects and smokers, suggesting that value-based decision-making can be illustrated by QRL at both the behavioural and neural levels.

Originating from early behavioural psychology, reinforcement learning is now a widely used approach in the fields of machine learning[1] and decision psychology[2]. It typically formalizes how one agent (a computer or animal) should take actions in unknown probabilistic environments to maximize its total reward. The agent selects actions according to value functions that describe the expectations for alternative decisions (value-based decision-making) and are influenced by reward, penalty and its beliefs about the current situation.

Recently, quantum computation techniques were successfully applied in the field of machine learning[3]. Based on the quantum superposition principle and quantum parallelism, quantum reinforcement learning (QRL) was proposed[4], combining quantum theory and reinforcement learning. This approach was later applied to robot navigation[5,6] and quantum machine learning[7,8]. Computer simulations[4,5] showed that QRL performs better on a large search space, learns faster and balances better between exploration and exploitation compared with classical reinforcement learning (CRL). It was shown theoretically that QRL can achieve quadratic improvements in learning efficiency and exponential improvements in performance for a broad class of learning problems[8]. This algorithm was also generalized to better adjust weights on favourable actions[6], which further demonstrated the robustness of this framework.

In a similar way to the introduction of quantum-inspired techniques in machine learning, quantum-inspired frameworks have also been introduced in psychology. There is evidence supporting quantum models for human behaviour: in the last decade, several cognitive scientists found that some behavioural paradoxes and effects (for example, the conjunction fallacy and the order effect) that resist explanations from classical probability theory could be explained well by quantum probability theory[9–16]. For example, one work showed the superiority of a quantum random walk model over classical Markov random walk models for a modified random-dot motion direction discrimination task and revealed quantum-like aspects of perceptual decisions[12]. Another study proved the superiority of one quantum model over the standard prospect model for a two-stage gambling task using the Bayesian factor model comparison[11]. In these works, the mathematical structure of quantum probability theory was emphasized, rather than the physical explanations for quantum mechanics. The assumption of low-level quantum physics processes is not necessary in these cases.

Although these quantum models were successful for some kinds of human decision-making, value-based decision-making[17] (the central issue of decision neuroscience and neuroeconomics[18–20] and a hallmark of high-level cognition) has not yet been tackled using quantum frameworks. Over the past decades, neuroscientific and psychological research on value-based decision-making has converged on the standard model of CRL[2,21,22], providing highly fruitful explanations and surprisingly accurate predictions. Due to some seemingly quantum-like features of value-based decision-making (in real life, making a choice per se can influence one's subjective values for the alternatives) and the theoretical efficiency of quantum learning algorithms, we aimed to test whether the QRL framework could also be useful and insightful for modelling value-based decision-making.

Moreover, previous quantum models for human behaviour were mainly supported by behavioural data. Only a few of them were supported by evidence from electroencephalography (EEG) analysis, lacking spatial resolution and locality information[23]. We have little knowledge about how these quantum-like mechanisms are

[1]Eye Center, Dept. of Ophthalmology, the First Affiliated Hospital of USTC, Hefei National Laboratory for Physical Sciences at the Microscale, School of Life Sciences, Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei, China. [2]Department of Statistics and Finance, School of Management, University of Science and Technology of China, Hefei, China. [3]School of Engineering and Information Technology, University of New South Wales, Canberra, Australian Capital Territory, Australia. [4]Shanghai Key Laboratory of Psychotic Disorders, Shanghai Mental Health Centre, Shanghai Jiao Tong University School of Medicine, Shanghai, China. [5]The First Affiliated Hospital of USTC, Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei, China. [6]Key Laboratory of Applied Brain and Cognitive Sciences, School of Business and Management, Shanghai International Studies University, Shanghai, China. [7]Theoretical Quantum Physics Laboratory, RIKEN Cluster for Pioneering Research, Wakoshi, Japan. [8]Department of Physics, The University of Michigan, Ann Arbor, MI, USA. [9]Hefei Medical Research Centre on Alcohol Addiction, Anhui Mental Health Centre, Hefei, China. [10]Academy of Psychology and Behaviour, Tianjin Normal University, Tianjin, China. [11]Centres for Biomedical Engineering, University of Science and Technology of China, Hefei, China. *e-mail: zxcustc@ustc.edu.cn

implemented in the brain (for example, which brain area might participate in quantum-like processes).

Our goal is to explore quantum-like aspects of processes underlying value-based decision-making. To this end, we applied QRL to value-based decision-making and studied how this quantum framework could be implemented at both the behavioural and the neural levels.

First, we collected behavioural and functional magnetic resonance imaging (fMRI) data from 58 healthy subjects (control group) and 43 nicotine addicts (smoking group) performing the Iowa Gambling Task (IGT)[24]. This is a value-based decision-making task that is well known among psychiatrists and neuroscientists, designed to evaluate the degree of decision-making defects. Its complexity and high ecological validity endow it with the ability to capture important components (motivational, learning and choice processes) underlying real-life decision-making. In previous works, many CRL models for the IGT were developed[25] to break down single task performance into these components. Some of these models were very successful in illustrating the subprocesses of value-based decision-making and explaining the behavioural differences between healthy people and patients with decision-making defects, such as drug addicts and brain-damaged populations. The smoking group was chosen based on the following considerations: smokers were reported to show impaired decision-making in the IGT[26,27]; the results from the smoking group could further validate the conclusions about the mechanisms of decision-making in healthy people; and, by means of behavioural modelling and model-based fMRI analysis, the mechanisms of addiction disorders were better understood[25].

Second, we developed two additional QRL models (quantum superposition state learning (QSL) and quantum superposition state plus perseverance (QSPP)) for the IGT and compared them to several well-established CRL models. We found that, in all groups, the QSPP model provided goodness-of-fit and simulation performance comparable to the best CRL models. We further replicated these results in 504 healthy subjects (super group) from an online IGT data pool[28].

Third, we executed fMRI analyses to compare the neural substrates relevant to the learning processes in the QSPP model and the best-fitting CRL model. Because learning involves the updating of outcome-evaluation-based internal states[29], we proposed and analysed several internal-state-related variables (generalized quantum distance and uncertainty-related interaction in the QSPP model) that are important in the QRL models and found that they were represented in the medial frontal gyrus (MeFG).

Altogether, our findings support the idea of quantum-like processes during value-based decision-making at both the behavioural and neural levels and provide the fMRI evidence connecting quantum cognition to neuroscience.

Results

Task design and subject performance. Healthy (control group) and smoking (smoking group) subjects were recruited to perform the IGT[24] in an fMRI scanner (Fig. 1a). In this task, the subjects selected one of four decks in each trial to obtain potential reward or penalty points. They were required to learn the goodness of each deck to maximize their total rewards. Regardless of which group they were in, subjects chose the good decks significantly more often in the last 20 trials than in the first 20 trials (control group (two-tailed paired t test): $t(57)=17.84$, $P<0.001$, Cohen's $d=3.52$, 95% confidence interval: 0.54, 0.68; smoking group: $t(42)=10.02$, $P<0.001$, Cohen's $d=2.20$, 95% confidence interval: 0.39, 0.60), meaning that most of them gradually learned the task (Fig. 1b,c). In addition to these two groups, we also analysed behavioural data from an online IGT data pool (super group of 504 healthy subjects; see also Supplementary Fig. 1).

The QSPP model performed well. After a careful search of the literature, we implemented 12 CRL models designed to break down the IGT performance into subcomponents. Roughly speaking, all of them contained three stages[30] (Fig. 2a; see also Supplementary Fig. 2):

1. Evaluate the outcome (for example, the value plus perseverance (VPP) rule).
2. Update the expectancy (for example, the Decay learning rule).
3. Make a choice (for example, the trial-independent choice (TIC) rule).

Most of these models were developed in past works[30–32], some of which were very successful in understanding the mechanisms and defects of decision-making[32–35].

Using the principles of QRL, we developed two additional QRL models (Fig. 2b; see also Supplementary Fig. 2), the QSL model and the QSPP model, by replacing each classical component in the CRL models with its quantum counterpart (see also Supplementary Table 1):

1. The CRL models use the set-theoretic representation a_1, a_2, a_3, a_4 to describe four actions (selecting each deck), while the QRL models use the geometric representation $|a_1\rangle$, $|a_2\rangle$, $|a_3\rangle$, $|a_4\rangle$, with the four eigenvectors spanning an action space (Hilbert space).
2. The CRL models use the classical probability $p_j(t)$ to describe the tendency of selecting action a_j, while the QRL models use the quantum probability amplitude ψ_j such that the internal state can be represented as a superposition state in the action space, $|\psi\rangle=\sum_{j=1}^{4}\psi_j|a_j\rangle$.
3. The CRL models evaluate actions explicitly by value functions, while the QRL models evaluate actions implicitly by the probability amplitude.
4. The CRL models learn value functions from outcomes using various learning rules, while the QRL models learn the probability amplitude from experience using an amplitude-updating algorithm (Grover iteration, $\hat{U}_G$).
5. The CRL models generate actions according to various choice rules, while the QRL models generate actions by superposition state collapse. Therefore, the QRL models work in a quantum-like manner different from the CRL models.

We proposed a geometric and algebraic explanation for the main operator $\hat{U}_G$ in the QRL framework (Fig. 2c; see also Supplementary Figs. 3 and 4 and the Further explanations of the quantum operator $\hat{U}_G$ section in the Supplementary Methods). On the Bloch sphere, in each trial, the north pole $\hat{\mathbf{z}}=|a\rangle$ represents the chosen action and the south pole $|a_\perp\rangle$ represents the unchosen actions. The vector $\hat{\mathbf{z}}'=|\psi\rangle$ is the internal superposition state, while its closeness to the poles reflects the weights on actions. If $|\psi\rangle$ moves to the north pole, the agent will choose the same action in the next trial, but if $|\psi\rangle$ is rotated to the south pole, the agent will never choose the same action in the next trial. During learning, the main operator $\hat{U}_G=\hat{Q}_2\hat{Q}_1$ takes a two-step rotation: $\hat{Q}_1$ first rotates $|\psi\rangle$ into $|\psi'\rangle$ around $\hat{\mathbf{z}}=|a\rangle$ (blue circle) and then $\hat{Q}_2$ rotates $|\psi'\rangle$ into $|\psi''\rangle$ around $\hat{\mathbf{z}}'=|\psi\rangle$ (purple circle), changing the closeness of the state vector to the two poles. The rotation angles are computed from the outcome (a linear mapping from utility to rotation angles).

For simplicity, the VPP models (models with the VPP rule) and the QSPP model are considered to be two-term models, while the others are considered to be one-term models, based on whether the perseverance part is included.

These models were fitted using the optimization algorithm to maximize the log likelihood (LL) of each subject's choice sequence. To test the feasibility of the QRL models, we performed model comparisons based on the goodness-of-fit criterion and the

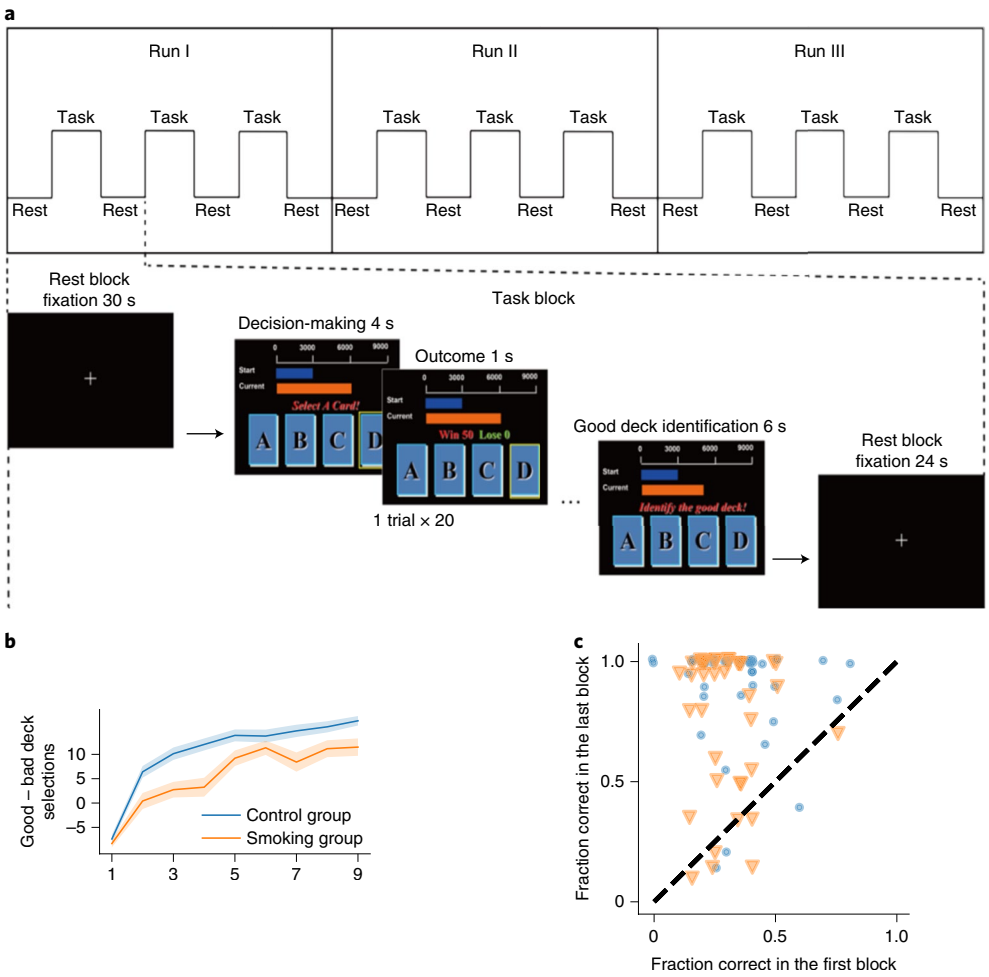

Fig. 1 | Task diagram and task performance. a, An IGT diagram and fMRI scan processes. There were three 7-min fMRI scan runs, separated by intervals of about 1 min. Each scan consisted of one 30-s rest block and three 106-s task and 24-s rest cycles. The last 6 s of each task block were designed for good deck identification, which was not analysed in this study. The first 100 s contained 20 trials and each trial was divided into two events: during the decision-making phase, 4 s were provided for card selection, and a random selection was made if no decision had been made during this period; during the outcome phase, the reward and penalty were presented on the screen for 1 s. There were no inter-trial intervals. The blue bar above the decks showed the initial 3,000 points throughout the task and the orange bar showed the current accumulated points. **b**, The proportion of good minus bad deck selections in the 20-trial blocks. The shaded regions indicate the standard error. **c**, Subjects chose the good decks significantly more often in the last 20 trials than in the first 20 trials. The markers are slightly jittered to show all subjects (blue circle, control group; orange triangle, smoking group). The dashed diagonal line represents the equality line.

simulation method[30] (see also Supplementary Results, section II for statistical test results).

For the goodness-of-fit criterion, the corrected Akaike's Information Criterion (AICc)[36] and the Bayesian Information Criterion (BIC)[37] provide a direct assessment of one-step-ahead predictions. Smaller AICc or BIC values represent better fits. The mean AICc values for all of the models in each group are presented in Fig. 3a–c. Here all models outperformed the baseline model, which assumes that the agent chooses each deck at a fixed probability. The QSL model provided better fits than all of the one-term CRL models (red bars) in the smoking and super groups and was comparable to the best one-term CRL PVLDecayTDC model in the control group. The QSPP model provided better fits in all three groups than all other CRL models, including the VPPDeltaTIC model, which was reported to be the best-fitting model among CRL models in former works[32]. The BIC values for all models provide similar results (Fig. 3d–f). In addition to frequentist model comparison, we further applied Bayesian model comparison based on the variational Bayesian method, treating the model as a random variable and estimating the distribution over model space[38]. Based on the AICc and BIC, we obtained the expected model likelihood, producing consistent results (Fig. 4). The QSPP model also showed an inferred frequency larger than any other CRL models. The two QRL models together provided an exceedance probability larger than 0.99 in all cases.

Unlike the goodness-of-fit criterion, the simulation method[30] was designed to assess the accuracy of one model generating predictions for entire choice sequences according to model parameters, without relying on the subjects' choice history. Based on the simulated sequences, we computed the mean square errors (MSEs) of the proportion of choice (PoC; Fig. 5a–c) and the MSEs of each deck selection (EDS; Fig. 5d–f). Here all the QRL models (orange bars) had a performance comparable to (or even better than) the best CRL models (red or blue bars) in all groups. Additional checks of the above distances between the behaviour of subjects and models are presented in Supplementary Fig. 5.

In conclusion, all comparison results indicate that the QSPP model is comparable to (or even better than) the CRL models at

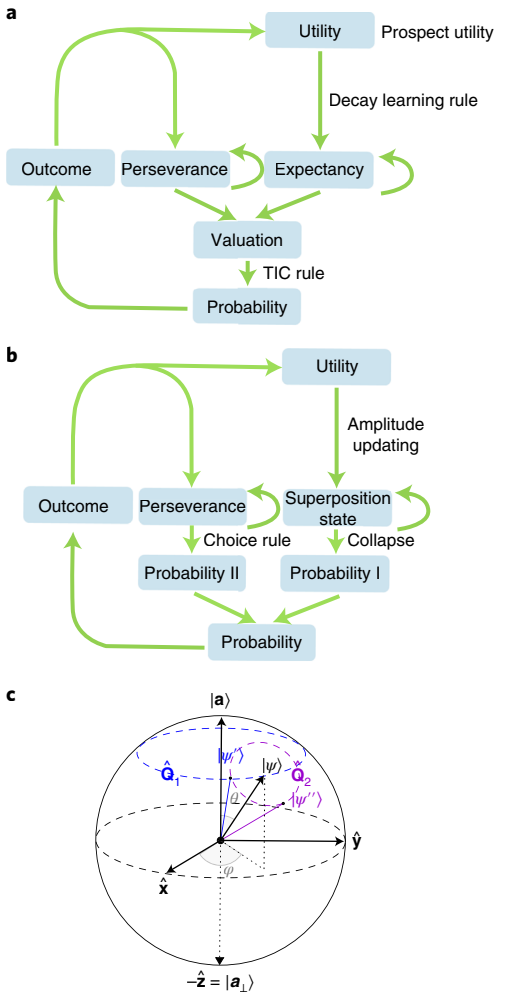

Fig. 2 | Diagrams of model architecture. a, The VPPDecayTIC model. VPPDecayTIC[31] is a CRL model containing three stages: (1) evaluate the outcome with the VPP rule; (2) update the expectancy with the Decay learning rule; and (3) make a choice with the TIC rule. **b**, The QSPP model. QSPP is a QRL model; the perseverance part is absent in one-term CRL models and in the one-term QRL model (QSL). Model details and comparisons can be found in the main text. **c**, A geometric explanation of $\hat{U}_G$ that adjusts the weights of the chosen action and unchosen actions. The north pole $\hat{\mathbf{z}} = |a\rangle$ on the surface of the unit sphere is the chosen action and the south pole $|a_\perp\rangle$ is the unchosen one. The vector $\hat{\mathbf{z}}' = |\psi\rangle$ is the internal state, while its closeness to the poles reflects the weights on the actions. The spherical coordinate parameters (polar angle θ and azimuthal angle φ) and the rectangular coordinate axes ($\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$) are determined by the Bloch sphere representation for our QRL models. During learning, the main operator $\hat{U}_G$ first rotates $|\psi\rangle$ into $|\psi'\rangle$ around $\hat{\mathbf{z}} = |a\rangle$ (blue circle) and then rotates $|\psi'\rangle$ into $|\psi''\rangle$ around $\hat{\mathbf{z}}' = |\psi\rangle$ (purple circle), changing the closeness of the state vector to the two poles (for details, see the section "Further explanations of the quantum operator $\hat{U}_G$" in the Supplementary Methods).

the behavioural level, proving the QRL to be a powerful framework to describe value-based decision-making behaviour. For the following analyses, we mainly concentrated on our QSPP model and the best-fitting VPPDecayTIC model as the representatives of the QRL and CRL models, respectively. We analysed the similarity in their parameters, showing that the VPPDecayTIC model was a reasonable counterpart of the QSPP model (see Supplementary Results, section III).

Generalized quantum distances were represented at the neural level. To illustrate the learning process formalized by the QSPP model, we factorized the main unitary operator $\hat{U}_G$ and acquired one of the most important properties of this unitary operator: quantum transition amplitudes (see the Quantum transition amplitudes and probabilities section in the Supplementary Methods). Just as electrons can jump between energy levels when absorbing or releasing energy, the agent's preference can change between the unchosen actions and the chosen action when acquiring information from the environment (learning from outcomes). Another way to depict the learning-induced change is the quantum distance between superposition states in two adjacent trials. Such distances are pervasive and meaningful in the field of quantum information[39], providing an approach to evaluate the distinguishability of different states (how close two states are). In our case, this distance evaluates to what degree the state is influenced by learning through unitary iteration (see the Quantum distances section in the Supplementary Methods).

Although the transition amplitude and the quantum distance have different definitions, meanings and concrete values, we found that there is a mathematical duality between the transition amplitude and the quantum distance (see the Relations between quantum transition amplitudes and quantum distances section in the Supplementary Methods), similar to the two sides of a coin. Therefore, we defined the generalized quantum distance as the geometric average of these two signals, catching the overall effect of $\hat{U}_G$. We computed the generalized quantum distance for each trial (Supplementary Fig. 6) and found that, on average, there was a downtrend over time, which means that the transition of preferences and the change of states induced by outcome learning are reduced gradually by accumulated knowledge about the task rule.

We then executed model-based fMRI analysis[40], which can relate model predictions to fMRI data to locate hypothesized decision-making processes in the brain and discriminate competing hypotheses. We analysed the generalized quantum distance based on the blood-oxygen-level dependent (BOLD) data at the outcome period in two groups (see generalized linear model (GLM) 3 in the fMRI data analyses section in the Methods). All reported voxels survived the family-wise error correction at a cluster level threshold of $P < 0.05$, with $P_{\text{uncorrected}} < 0.001$.

In the control group, the generalized quantum distance significantly activated the MeFG/anterior cingulate cortex (ACC), the precentral gyrus, the insula, the precuneus, the left fusiform gyrus, the inferior parietal lobule (IPL), the middle frontal gyrus (MiFG) and the right inferior temporal gyrus (ITG)/fusiform gyrus (Fig. 6; see also Supplementary Table 2), showing a unique quantum-like neural mechanism for how the internal state is changed due to external information.

In addition, we did not find activation related to generalized quantum distances in the smoking group, indicating that their representation might be impaired in smokers.

Neural substrates related to uncertainty revealed by the QSPP model. Uncertainty might be another useful variable to study to better understand the differences between the learning processes reflected by the VPPDecayTIC and QSPP models. Uncertainty about internal states, an important task-related variable represented at the neural level[41,42], reflects the external unstable environment, interacts with outcomes[43] and assists learning[44,45]. We computed the uncertainty provided by the two models (see the Uncertainty section in the Supplementary Methods). Uncertainty decreased over time (Supplementary Fig. 7), meaning that the subjects' uncertainty about the task rule dropped gradually. We then performed another model-based fMRI analysis, studying uncertainty and the interaction of uncertainty and penalty/reward based on the BOLD data (see GLM4–5 in the fMRI data analyses section in the Methods,

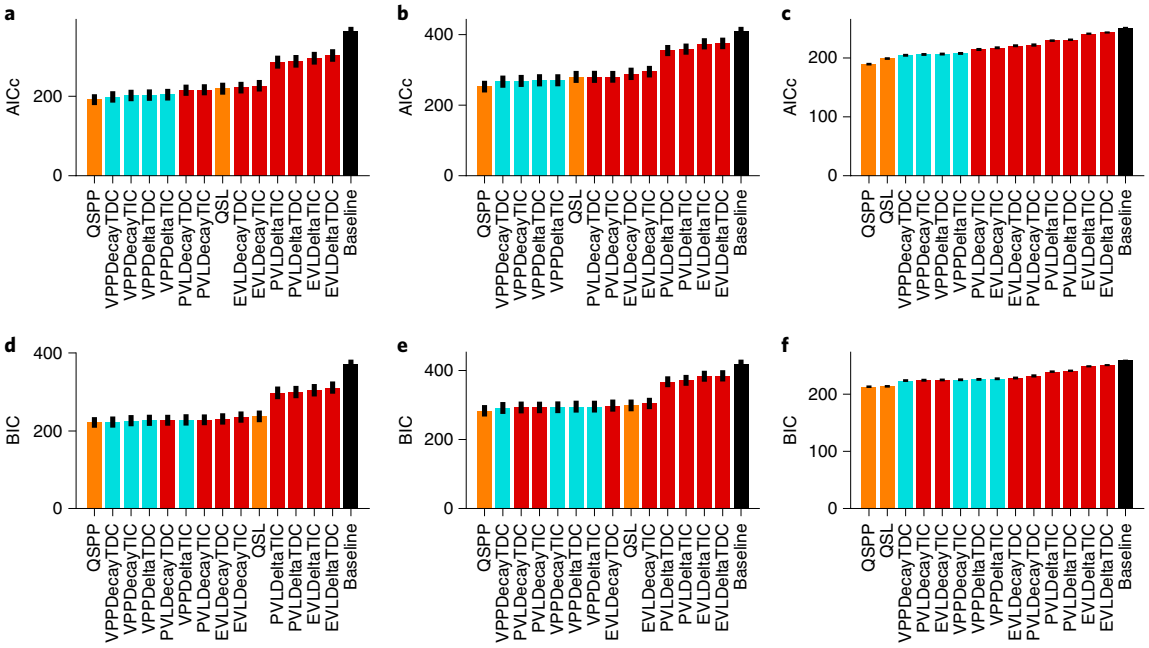

Fig. 3 | The AICc and BIC of each model, computed separately for each group. a, The AICc in the control group. **b**, The AICc in the smoking group. **c**, The AICc in the super group, with a different range from the first two because there were fewer trials for the super group. **d**–**f**, The BICs in the control (**d**), smoking (**e**) and super groups (**f**). The orange bars represent the QRL models, including the QSL (one-term) and the QSPP (two-term) models. The light blue bars represent the two-term CRL models, while the red ones show the one-term CRL models. All panels demonstrate the good performance of the QRL models. The error bars indicate the standard error.

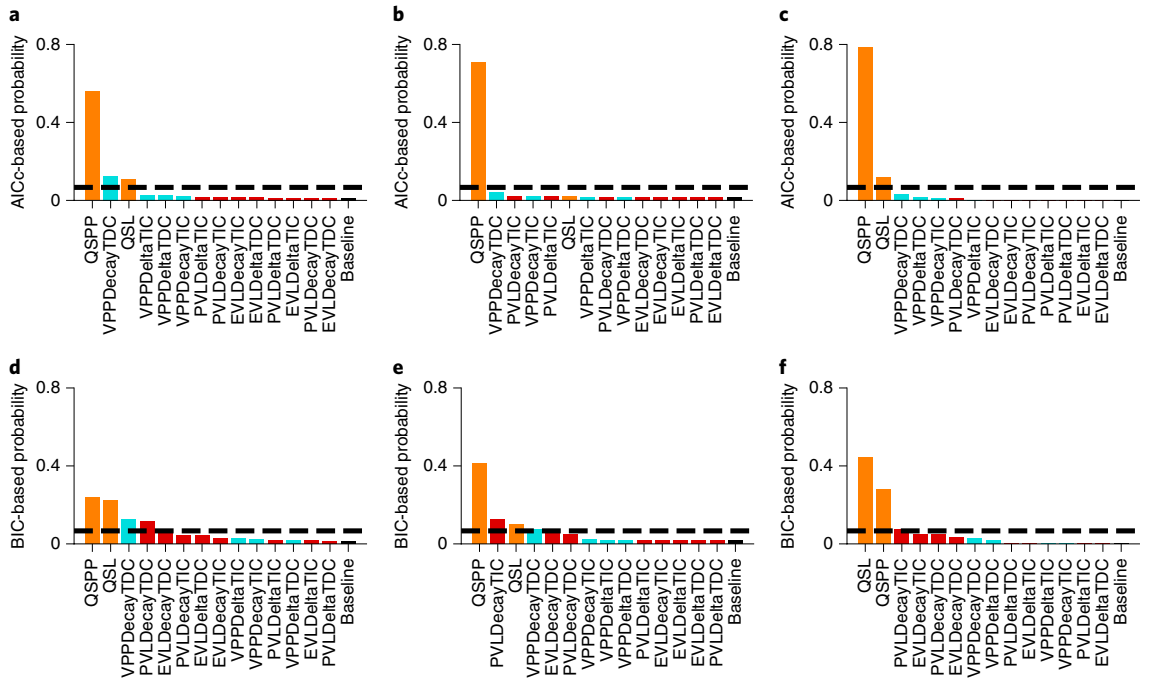

Fig. 4 | The inferred model probability of each model, computed separately for each group. a–**c**, The AICc-based inferred probability in the control (**a**), smoking (**b**) and super groups (**c**). **d**–**f**, The BIC-based inferred probability in the control (**d**), smoking (**e**) and super groups (**f**). The orange bars represent the QRL models. The light blue bars represent the two-term CRL models, while the red ones show the one-term CRL models. All panels demonstrate the good performance of the QRL models; the dotted line shows the average level.

fMRI results of uncertainty in Supplementary Results, section VI and Supplementary Table 3). Penalty and reward were separated because they were reported to have different roles in IGT processing and performance[46].

In the control group, the difference (control QSPP − control VPPDecayTIC) in the effects of the uncertainty × penalty interaction showed significant activation in the right MeFG, the left orbital frontal cortex (OFC), the left MeFG/ACC, the middle temporal

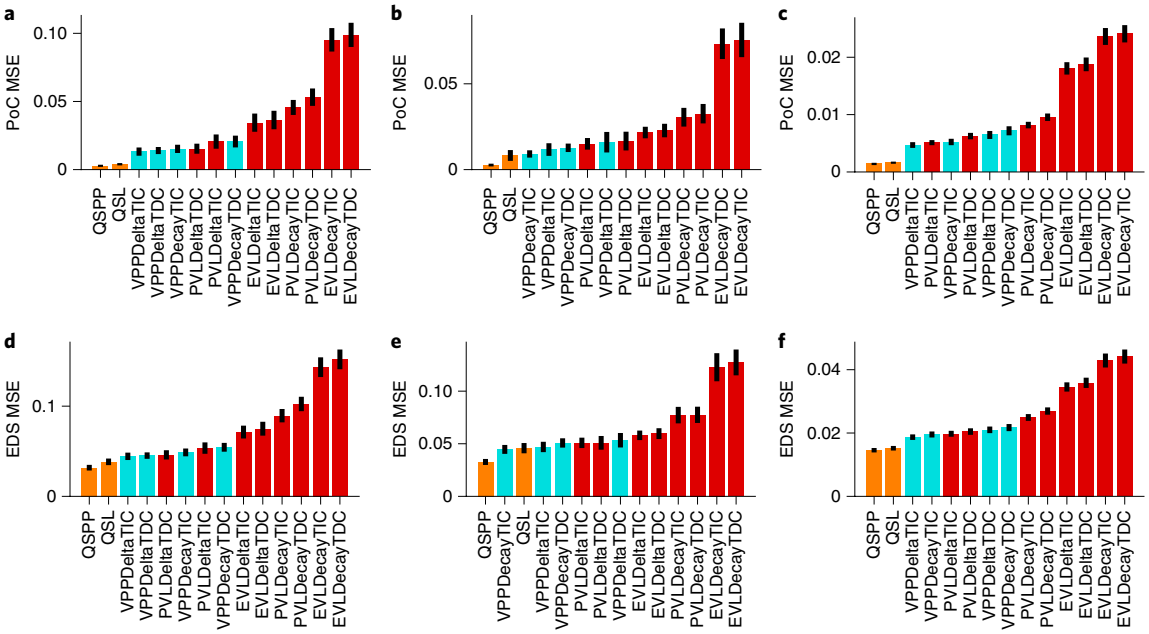

Fig. 5 | The simulation results of each model, computed separately for each group. a–c, The MSEs of the PoC in the control (**a**), smoking (**b**) and super groups (**c**). **d–f**, The MSEs of EDS in the control (**d**), smoking (**e**) and super groups (**f**). All panels demonstrate the good performance of the QRL models. All error bars indicate the standard error.

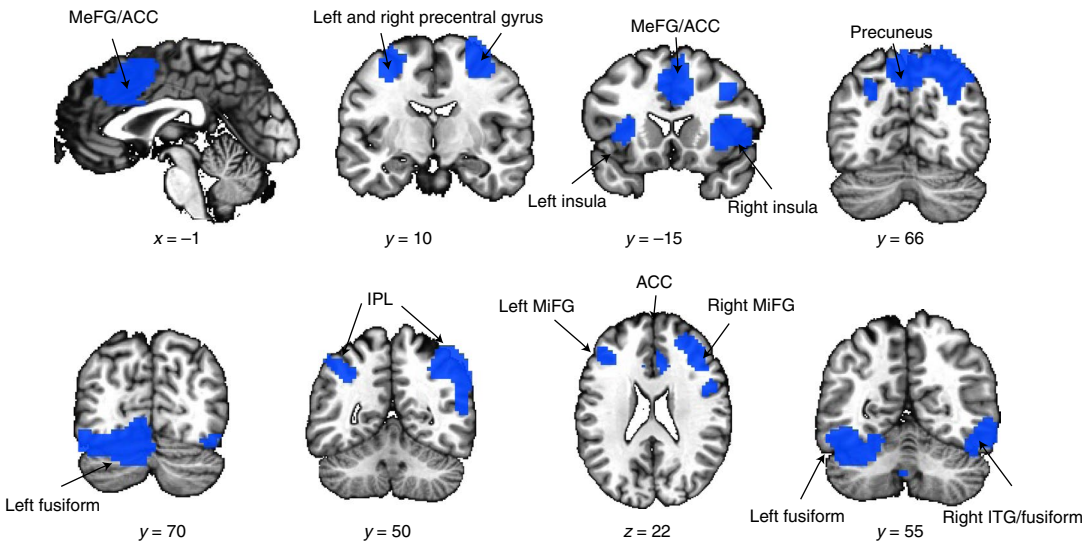

Fig. 6 | Generalized quantum distance (computed by the QSPP model)-related activity in the control group. The generalized quantum distance significantly activated the MeFG/ACC, the precentral gyrus, the insula, the precuneus, the left fusiform gyrus, the IPL, the MiFG and the right ITG/fusiform gyrus. All coordinates (in mm) are plotted in RAI (right–anterior–inferior, also known as radiological coordinates) order (i.e., −right, +left; −anterior, +posterior; −inferior, +superior).

gyrus (MiTG) and the left MiTG/angular gyrus/posterior cingulate gyrus (PCC)/precuneus (Fig. 7a,b; see also Supplementary Table 4). These results revealed distinct quantum-like mechanisms for how learning from a penalty is modulated by uncertainty.

In the smoking group, we also found the difference (smoking QSPP − smoking VPPDecayTIC) of effects of the interaction positively related to the left MeFG, left MiTG and left IPL (Fig. 7a,c). The other areas activated in the control group showed no activation in the smoking group.

For the interaction of uncertainty and reward, we found activation in the left MeFG (nearly significant), left MiTG and left MiTG/angular gyrus in the control group (Fig. 7d; see also Supplementary Table 5). No significant difference was found between the two models related to the interaction of uncertainty and reward in the smoking group.

Discussion

In the present study, we developed two additional QRL models for the IGT. We compared these two QRL models with 12 CRL models and found that the QSPP model was comparable to the best CRL models. We also studied the generalized quantum distances and the uncertainty × penalty/reward interaction based on fMRI data

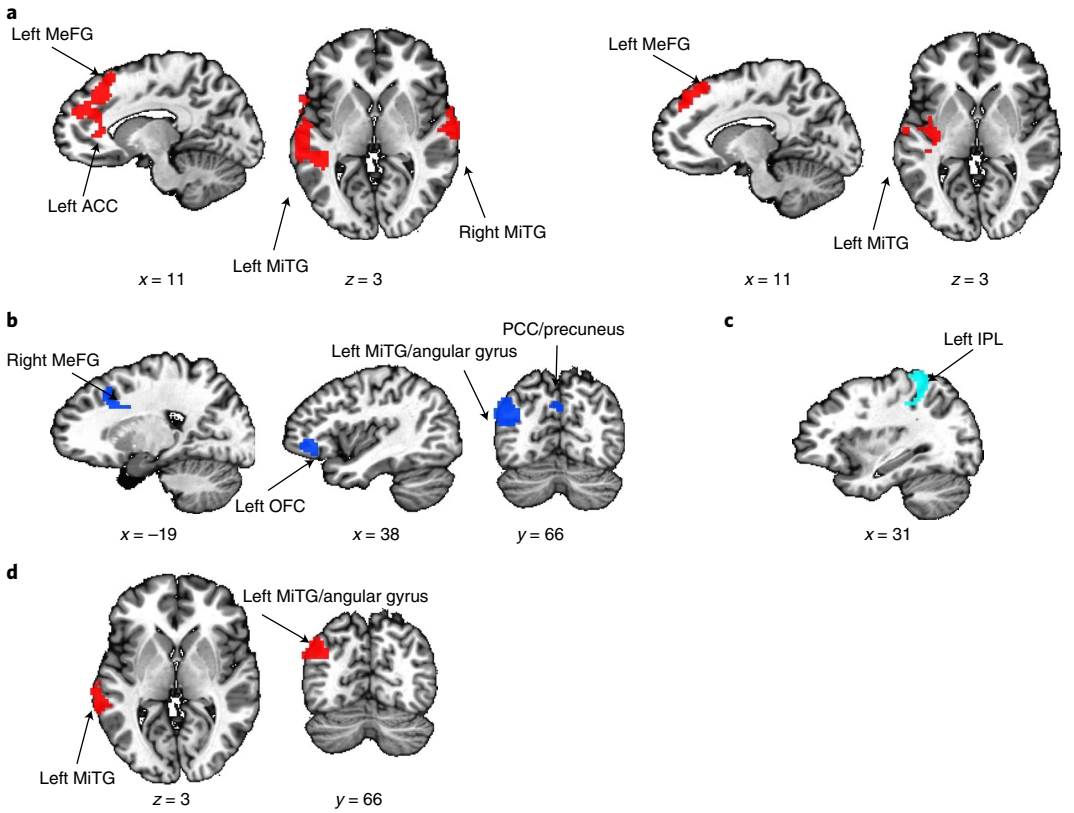

Fig. 7 | fMRI results of the uncertainty × penalty/reward interaction. a, The uncertainty × penalty interaction positively related to the left MeFG and left MiTG in both the control group (left panel; control QSPP − control VPPDecayTIC) and the smoking group (right panel; smoking QSPP − smoking VPPDecayTIC). **b**, The uncertainty × penalty interaction positively related to the right MeFG, left OFC and left MiTG/angular gyrus/PCC/precuneus in the control group. **c**, The uncertainty × penalty interaction positively related to the left IPL in the smoking group. **d**, The uncertainty × reward interaction positively related to the MiTG and the left MiTG/angular gyrus in the control group (control QSPP − control VPPDecayTIC). All coordinates are plotted (in mm) in RAI order (i.e., −right, +left; −anterior, +posterior; −inferior, +superior).

in healthy subjects and smokers. We found several consistencies and differences in brain activity between the control and smoking groups, which reveal the important brain regions for QRL; they are discussed below.

To compare these models, we used the goodness-of-fit criterion and the simulation method. The goodness-of-fit criterion evaluated the accuracy of the one-step-ahead predictions of the models. Using the AICc and BIC, the QSPP model provided a better fit than all CRL models. Unlike one-step-ahead predictions, the simulation method produced whole selection sequences without relying on the subjects' real choices. With two MSE indexes, the QSPP model kept its good performance. More importantly, the QSPP model showed superior performance in all three groups, indicating that this can be generalized for more diverse subjects.

The superiority of the one-term Decay models over the one-term Delta models using the goodness-of-fit criterion was reversed with the simulation method (see Supplementary Results, section I). This phenomenon was also reported in previous works; this could be because the two rules have different reliances on past outcomes and past choices[30,47]. The Delta rule provides better long-term (whole sequence) predictions and the Decay rule makes better short-term (one-step-ahead) predictions. The QRL models achieved good performance in both the goodness-of-fit criterion and the simulation method, indicating that they make good predictions for both the long term and short term.

Generally, the learning process in value-based decision-making is the adjustment of the weights on each action (for example, the preference/tendency for each deck). In the CRL models, the weights are implemented by value functions (the expectancy for decks), while in the QRL models, the weights are expressed by the probability amplitude for each action (superposition state overall). Cognitively, during each trial's learning period in the CRL models, the weight on the chosen action is updated (for example, by prediction errors) and the weights on unchosen actions decay or remain unchanged, independently of the chosen action. However, in the QRL models, the superposition state can be visualized as a point (vector) on the surface of a three-dimensional sphere (Bloch sphere) and the adjustment of weights on each action is simultaneously reflected in the movement of the point location, which is implemented as the rotation of the point (superposition state) on the sphere surface around an axis. In short, the CRL models we used adjust the weights on the chosen action and unchosen actions independently, separably and locally during the valuation process, while our QRL models adjust the weights on the chosen action and unchosen actions dependently, inseparably and globally.

In contrast to simple modifications of the CRL models (for example, adding or deleting components in the CRL models), our QRL models provide us with a different perspective from traditional ones and point out other possible directions for modelling, such as geometric state representation, unitary state evolution and Grover iteration updating.

The QRL models are also different from the original QRL framework[4,6], in which the cumulative value function (such as the expectancy for decks in the IGT, the state value function $V(s)$ and the state-action value function $Q(s, a)$ in reinforcement learning) was used for superposition state updating. In practice, we found that the

incorporation of cumulative value functions did not significantly increase model performance. In the field of decision neuroscience, all reinforcement learning models involve the calculation of cumulative values, while our QRL models do not assume the use of such a component. It seems to be consistent with our daily life experiences: we sometimes do not feel the existence of such a computational process during this kind of decision-making. Although we can explain that this process is automatic, habitual and unconscious, our QRL models provide a potential different way: cumulative value computation could be absent during value-based decision-making. Whether cumulative value functions are profitable in the QRL models is an open question. In future research, it is worth investigating whether they can further improve the performance.

Although the QRL models did not overwhelmingly outperform the CRL models (for example, some behavioural effects can be explained by only the QRL models), one should be aware that the QRL framework is still in its infancy and undoubtedly deserves additional studies. Compared with the classical framework, the QRL framework has several potential theoretical advantages:

1. Quantum probability theory, as a generalized probability theory, provides a stronger representation for tasks and internal states (such as allowing the internal states to be in an indefinite state before an action is taken[10]).
2. From the perspective of information processing, there are ample quantum features (such as coherence and entanglement) not (fully) utilized in the current models that could be beneficial in the future.
3. In principle, the Grover algorithm can have a quadratic speed-up (i.e., the improvement of time complexity from $O(N)$ to $O(N^{1/2})$) on learning over classical algorithms[48]. Recently, several theoretical works[8,49] showed that quantum enhancement can be achieved in learning efficiency, learning performance and meta-learning using the QRL framework.

We also discuss the possible limitations of our QRL models, point out areas for improving them and call attention to potential future directions:

1. The general QRL algorithms consist of two sequential stages: mapping outcomes (reward or penalty) to learning factors and mapping learning factors to a parametric transformation (unitary operator). In our models, we used the utility function in the first stage and a two-step rotation on the Bloch sphere in the second stage, while a substantial number of function forms remain largely unexplored.
2. The learning process of our QRL algorithm (and of previous QRL algorithms) is described by an equivalent two-dimensional subspace (spanned by the chosen action vector $|a\rangle$ and the unchosen action vector $|a_\perp\rangle$) of the full four-dimensional space, enabling its visualization on the Bloch sphere. To show more detailed information during high-dimensional learning, new geometric visualization methods are required, such as the generalization of the Bloch sphere to higher dimensions[50].
3. There are many other advanced techniques in quantum cognition modelling[51] (such as mixed state, von Neumann entropy and positive operator valued measures) and developments in QRL algorithms (such as reinforcement learning using quantum Boltzmann machines[52]) that may be enlightening and useful.
4. We showed the superior performance of our models for only the IGT, while more decision-making tasks are worthy of future investigation. In addition, new value-based decision-making tasks should be designed to further explore the differences between the QRL and CRL models.

We think it is too early to conclude that we should abandon one framework and accept another one because they both remain largely undeveloped and have much potential. More importantly, the QRL framework reveals a different perspective on human behaviour and these two frameworks can reap a substantial number of benefits from each other.

This fMRI study supports the idea of quantum cognition at the neural level. For the learning process of unitary iteration in the QSPP model, the quantum transition amplitude describes the transition between actions due to external information. The quantum distance measures the distinguishability between two quantum states (that is, how close they are) and thus describes the influence on internal states by learning from outcomes. The generalized quantum distances take into consideration the duality between the quantum distance and the transition amplitude and thus catch the overall property of the main operator. The transition amplitude and the quantum distance showed activation consistent with the generalized quantum distance, indicating the rationality of the definition of our generalized quantum distance (see Supplementary Results, section V). The generalized quantum distance can be considered as the interaction between the internal superposition state and the outcome, sharing a similar critical concept with the interaction of uncertainty and outcome (also discussed in the next paragraph). We found that the generalized quantum distances were represented in the MeFG/ACC, the precentral gyrus, the insula, the MiFG, the precuneus, the IPL and the fusiform gyrus (because the MeFG is also activated for the uncertainty × penalty/reward interaction, which we discuss later). The MiFG, the precuneus and the fusiform gyrus were reported to be activated in many decision-making tasks under uncertainty[53–56]. The IPL was reported to play a role in decision-making under uncertainty, updating the evaluation for actions[56,57]. The precentral gyrus is related to the preparation and execution of motor responses in decision-making[55]. The insula was implicated in interoception, emotion and risky decision-making, such as risk experience representation and prediction error representation[55,56,58]. Therefore, the activation in these brain areas reveals quantum-like aspects of state representation, learning and other possible downstream processes, such as evaluation and action generation related to internal states.

Uncertainty is an internal-state attribute that reflects the external environment; it is influenced by outcomes (penalty and reward). Therefore, similar to generalized quantum distance, the uncertainty × penalty/reward interaction revealed a unique quantum-like modulation effect of outcomes on the internal state, involving integration and learning. Consistent with previous results, we found the same brain area (the MeFG) activated again in both the control and the smoking groups, indicating that the quantum-like mechanisms in the MeFG may be the hub of quantum-like effects in valued-based decision-making, affecting well-known cognitive processes (for example, uncertainty influenced by penalty) during reinforcement learning. Our MeFG is a part of the medial frontal cortex, which was reported to be involved in decision-making[21,59–62], such as the generation of internal voluntary action selection guided by memory and endogenous cues. Therefore, the MeFG might participate in state representation and collapse (action generation) in a quantum-like manner.

We also found that the uncertainty × penalty interaction activated the ACC, the OFC, the MiTG, the angular gyrus/MiTG and the PCC/precuneus and that the uncertainty × reward interaction activated the MiTG and the angular gyrus/MiTG. The ACC was reported to represent prediction errors and learning rate[63] and monitor and integrate outcomes[44]. The OFC was reported to represent prediction errors, update values[64], evaluate the amount of uncertainty[65,66], optimize learning rates under uncertainty[54] and represent abstract features of external outcomes[63]. The angular gyrus has been implicated in risky decision-making requiring visual attention[67]. The PCC was reported to have functional connections with the MeFG, the precuneus, the MiFG, the fusiform gyrus and the

precentral gyrus during the IGT[56], which were activated by generalized quantum distances. These areas activated by the interaction might reflect a role in monitoring learning processes and connecting relevant brain areas and functions. Altogether, these areas activated by internal-state-related variables, including the MeFG hub, might work as a network in a quantum-like manner, representing internal states, learning from the outcomes and generating actions.

The MeFG and MiTG were also activated by uncertainty × penalty interaction in the smoking group, supporting the generalization of our fMRI results. By contrast, the ACC, the OFC, the angular gyrus/MiTG and the PCC/precuneus showed quantum-like interaction-related activity in the control group but not in the smoking group, meaning that they are more sensitive to smoking addiction. This phenomenon indicates that the decreased decision-making ability in smokers might relate to the inability of these regions to represent the interaction, although further work is warranted to determine the exact causation. Combining these results and the discussion in the last paragraph indicates that the smokers were still able to represent the internal states and generate actions in a quantum-like way, but show an injured ability to represent outcomes and learn from experience.

In addition, the uncertainty × reward interaction showed only nearly significant activation in the MeFG between the two models, which might be because people are usually more sensitive to penalty than reward[68]. Moreover, the IPL was activated by the uncertainty × penalty interaction in the smoking group but not in the control group; this requires further studies. This activation in the IPL also indicates that the reduced activation in other areas in the smoking group is not due to the number of subjects or our data quality, but reflects the effect of addiction.

Although quantum frameworks are successful in modelling cognition, how quantum-like mechanisms are feasible in the brain is still unclear[69]. Of course, unlike other quantum phenomena in biology[70], our fMRI results do not suggest that our brain is quantized. Even a classical brain can generate quantum-like functions and behaviour following quantum principles. Many scientists believe that quantum probability theory is usually more emphasized in quantum cognition than its physical counterpart[71]. Physicists showed that the classical dynamic system might have quantum-like behaviour under coarse-grained measurements[72–75] and that an artificial neural network can approximate the wave function of a system efficiently[76]. There was also an existence proof that the common quantum probability operators can be implemented in classical recurrent neural networks with neural oscillation and synchronization[77]. Based on these and our work, we could imagine a possible approach bridging quantum cognition and neural implementation (see Supplementary Discussion, section II).

In conclusion, we have shown that the QRL framework can be useful in the fields of decision neuroscience and neuroeconomics and that it is a potential competitor of CRL. In view of the great success achieved by CRL in studies of emotion[78], psychiatric disorders such as addiction and depression[79,80], social behaviour[81], free will[82] and many other cognitive functions, we hope that QRL will also shed light on them. It is necessary to perform wider and deeper studies on the QRL framework. Since QRL is one type of quantum learning[10], we also hope that other methods, including quantum Bayesian learning[83] and quantum neural network learning[84], will be explored in the future. Because of the marriage of QRL and decision neuroscience, the possible resulting discipline could be termed as quantum neuroeconomics. "Is there a quantum neuroeconomics?" asked Piotrowski and Sladkowski[85]. Now, we are approaching the answer.

Methods

Participants. Fifty-eight healthy subjects (control group: 9 females; mean age ± s.d.: 23.42 ± 2.36 yr; mean education ± s.d.: 16.58 ± 1.79 yr) and 43 smokers (smoking group: all men; mean age ± s.d.: 24.44 ± 2.34 yr; mean education ± s.d.: 16.13 ± 1.92 yr; more than 10 cigarettes per day for at least 1 yr) participated in the IGT and were included in the following behavioural modelling analysis. While no statistical methods were used to predetermine sample sizes, previous works on the IGT using fMRI used sample sizes typically between 15 and 30 (for example, see refs. [86–88]) and our study had a reasonable number of subjects to produce stable results.

All subjects, with no prior knowledge of the task, were recruited by internet advertisements. They had normal or corrected-to-normal vision and were right-handed. They did not have any major medical illnesses, major psychiatric disorders, neurological illnesses or a history of dependence on any drug (except nicotine in the smoking group), or gross structural abnormalities in their T1-weighted images. Informed consent was collected from all subjects and the study was authorized by the Research Ethics Committee of the University of Science and Technology of China and conformed to the tenets of the Declaration of Helsinki.

Procedure. Subjects were required to make repeated choices from four decks to gain or lose a certain number of points and were told that the final payment was determined by their final scores. Next, subjects completed the training session for the task outside the fMRI scanner; this lasted 5 min without any real payment. The session presented only equal pay-offs across all decks during the training session of the IGT to avoid disturbance. All the subjects were told that there were 'good decks' and 'bad decks' that would give rise to a net gain or net loss in the long term, respectively. Their purpose was to find out the invariable rule in the IGT through exploration from the beginning of the task and the training was just to become familiar with the operational interface. The difference between the training and experimental sessions was clear to the subjects. After a 10-min break, the subjects executed three in-scanner runs of the IGT. They had 3,000 initial task points as the 'start-up capital' and received a payment corresponding to the task points they obtained in the scanner. Every 100 points could be exchanged for 1 Chinese yuan (about USD 0.15). The 3,000 initial points were not included in behavioural analyses.

Data collection and analysis were not performed blind to the conditions of the experiments. There was no randomization in the data collection, the organization of the experimental conditions or stimulus presentations.

Task and stimuli. The IGT is exactly the same as in the previous studies[27,56]. In this task, there were four decks of cards (decks A, B, C and D, presented from left to right). On the front of each card there were gain points and possible loss points. The subject would obtain net points (gain − loss) in each trial.

Deck A gave 100 (gain) points for each card and −150, −200, −250, −300 and −350 (loss) points for five cards, respectively, out of every ten cards. Deck B also gave 100 (gain) points for each card and −1,250 (loss) points once out of every ten cards. On average, ten choices of deck A or deck B resulted in −250 net points. Deck C gave 50 (gain) points for every card and −25, −40, −50, −60 and −75 (loss) points for five cards, respectively, out of every ten cards. Deck D also gave 50 (gain) points for every card but −250 (loss) points once out of every ten cards. On average, ten choices of deck C or deck D resulted in 250 net points. Deck A and deck B were disadvantageous because they resulted in an overall loss, although they gave a relatively larger gain in most trials. Deck C and deck D were advantageous because they resulted in an overall gain, despite giving a relatively smaller gain in most trials. To encourage as many subjects as possible to find the rule, the task was extended to 180 trials from the original 100 trials.

The 180 trials were separated into three scan runs. Each scan run included three task blocks separated by a 24-s rest (fixation) block and each task block consisted of 20 trials. Each trial contained a 4-s decision period and a 1-s outcome period.

Online data pool. There was an online IGT data pool[28] of 617 healthy subjects from 10 studies. These subjects completed the IGT with 95–150 trials (95 trials: 15 subjects from 1 study; 100 trials: 504 subjects from 7 studies; 150 trials: 98 subjects from 2 studies). We analysed only the data from the 504 subjects, because the remaining 113 subjects performed IGT with a different number of trials and the group of 504 subjects could already represent the whole IGT data pool.

Computational modelling. *The baseline model.* This model serves as a baseline reference[89]. It assumes that the agent chooses each deck at a fixed probability. There are three free parameters here, p_A, p_B and p_C, for the probability of choosing A, B and C decks, respectively. $p_D = 1 - (p_A + p_B + p_C)$ is the remaining probability.

Expectancy valence learning (EVL) and prospect valence learning (PVL) models. The architecture of these two types of model can be found in Supplementary Fig. 2a.

Utility function. The expectancy utility function simply assumes that the utility is the weighted difference of current gain and loss[90]:

$$u(t) = (1 - W) \times \text{gain}(t) - W \times |\text{loss}(t)| \quad (1)$$

where W ($0<W<1$) measures the weights for loss points against gain points.

The prospect utility function[68] is believed to explain the gain–loss frequency effect[91]. The prospect utility $u(t)$ is:

$$u(t)=\begin{cases}x(t)^{\alpha} & \text{if} \quad x(t)\geq 0\\ -\lambda|x(t)|^{\alpha} & \text{if} \quad x(t)<0\end{cases} \tag{2}$$

where $x(t)=\text{gain}(t)-|\text{loss}(t)|$ is the net outcome in trial t, α ($0<\alpha<2$) is a shape parameter of the utility function and λ ($0<\lambda<10$) is a loss-aversion parameter. Raw pay-offs within data are divided by 100 (default scale)[92].

Learning rules. The Delta learning rule, or the Rescorla–Wagner rule[93], is widely used in the fields of reinforcement learning and IGT studies[89,90]:

$$E_j(t)=E_j(t-1)+k\delta_j(t-1)[u(t-1)-E_j(t-1)] \tag{3}$$

where $E_j(t)$ is the expectancy for deck j in trial t ($t\geq 2$ and $E_j(1)=0$) and $\delta_j(t)$ is 1 if deck j is chosen in trial t and 0 otherwise. Here k is the learning rate of the prediction error $[u(t-1)-E_j(t-1)]$.

The Decay learning rule[94] assumes that the expectancy about each deck will decay with time and the expectancy of the chosen deck in trial t is updated by $u(t)$:

$$E_j(t)=AE_j(t-1)+\delta_j(t-1)u(t-1) \tag{4}$$

where A ($0<A<1$) is a decay parameter describing the strength of expectancy discounting.

Choice rules. A choice rule can be described as a balance between exploration and exploitation. At first, agents explore the goodness of decks and after several trials they exploit the knowledge acquired before, to maximize their overall pay-off. A common choice rule is the Boltzmann exploration or SoftMax selection[89]:

$$\Pr[\delta_j(t)=1]=\frac{\exp\{\theta(t)E_j(t)\}}{\sum_{k=1}^{4}\exp\{\theta(t)E_k(t)\}} \tag{5}$$

where $\Pr[\bullet]$ is the probability function. The sensitivity parameter $\theta(t)$ can be determined by a trial-independent choice (TIC) rule:

$$\theta(t)=3^{c}-1 \tag{6}$$

where c ($0<c<3$) is a consistency parameter. A large c value indicates a deterministic choice and a small one suggests a random choice[47].

The sensitivity parameter $\theta(t)$ can also be computed from a trial-dependent choice (TDC) rule:

$$\theta(t)=(t/10)^{c} \tag{7}$$

where the consistency parameter c ($-5<c<5$) indicates gradually more deterministic choices if $c>0$ (for example, because of confidence) and gradually more random choices if $c<0$ (for example, because of boredom).

Model description. By combining these two utility functions (EVL and PVL), two learning rules (Delta and Decay) and two choice rules (TIC and TDC), we obtain the eight models[30]: EVLDeltaTDC, EVLDeltaTIC, EVLDecayTDC, EVLDecayTIC, PVLDeltaTDC, PVLDeltaTIC, PVLDecayTDC and PVLDecayTIC.

VPP models. The VPP model (VPPDeltaTIC) is an improved version of the PVLDeltaTIC model, incorporating the influence of perseverance[31].

The VPPDeltaTIC model assumes that the agent will take the perseverative strategy into consideration (Supplementary Fig. 2c). The perseverance strength $P_j(t)$ on deck j in trial t ($t\geq 2$ and $P_j(1)=0$) is defined as:

$$P_j(t)=\begin{cases}A_{\text{pers}}P_j(t-1) & \text{if} \quad \delta_j(t-1)=0\\ A_{\text{pers}}P_j(t-1)+\epsilon_{\text{p}} & \text{if} \quad \delta_j(t-1)=1 \quad \text{and} \quad x(t-1)\geq 0\\ A_{\text{pers}}P_j(t-1)+\epsilon_{\text{n}} & \text{if} \quad \delta_j(t-1)=1 \quad \text{and} \quad x(t-1)<0\end{cases} \tag{8}$$

where A_{pers} ($0<A_{\text{pers}}<1$) is a free decay parameter on the perseverance strengths of all decks. Also, ϵ_{p} ($-10<\epsilon_{\text{p}}<10$) is the influence of the net gain on perseverance and ϵ_{n} ($-10<\epsilon_{\text{n}}<10$) is the influence of the net loss on perseverance.

The value $V_j(t)$ is the weighted average of the expectancy $E_j(t)$ and the perseverance strength $P_j(t)$:

$$V_j(t)=wE_j(t)+(1-w)P_j(t) \tag{9}$$

where w is the reinforcement learning weight ($0<w<1$). A high w value indicates that the agent prefers reinforcement learning rather than perseverative strategy and vice versa. Note that $E_j(t)$ here is updated by the Delta learning rule for the VPPDeltaTIC model and by the Decay learning rule for the VPPDecayTIC model.

The choice probability here is also calculated by the SoftMax function but with $V_j(t)$:

$$\Pr[\delta_j(t)=1]=\frac{\exp\{\theta(t)V_j(t)\}}{\sum_{k=1}^{4}\exp\{\theta(t)V_k(t)\}} \tag{10}$$

By replacing the TIC rule with the TDC rule, we similarly obtain the VPPDecayTDC and the VPPDeltaTDC models.

The QSL model. The architecture of this model can be found in Supplementary Fig. 2b.

State representation and action selection. In the IGT, there are four actions representing the selection of each deck. In the classical models, they are modelled as elements of the action set. Thus, the probabilities of each action form a measure on the discrete action set, requiring the sum of all the probabilities to equal 1. If we consider $p_j(t)$ as the probability of choosing deck j in trial t, then the constraint is:

$$\sum_{j=1}^{4}p_j(t)=1 \tag{11}$$

The QSL model allows the agent to be in an indefinite state (action state), formally called the superposition state. The superposition state is a vector in a D-dimensional Hilbert space (action space) spanned by D orthogonal basis vectors denoted by the symbol $|a_j\rangle$, $j=1,\ldots,D$, if using the Dirac bra-ket notation. For the IGT, $D=4$ and these basis vectors, or eigenvectors, $|a_1\rangle$, $|a_2\rangle$, $|a_3\rangle$ and $|a_4\rangle$ represent the actions of choosing the A, B, C or D deck, respectively. Then the superposition state of the agent in trial t is:

$$|\psi(t)\rangle=\sum_{j=1}^{4}\psi_j|a_j\rangle \tag{12}$$

$$\psi_j=\langle a_j|\psi(t)\rangle \tag{13}$$

where ψ_j, the inner product of $|\psi(t)\rangle$ and $|a_j\rangle$, is the probability amplitude of each action, which can be a complex number.

During the decision period, the superposition state $|\psi(t)\rangle$ collapses onto one of the eigenvectors $|a_j\rangle$, with the probability of the squared magnitude of the corresponding amplitude. The collapse process is the so-called "action selection":

$$\Pr[\delta_j(t)=1]=|\psi_j|^2 \tag{14}$$

Thus we have the corresponding constraint:

$$\sum_{j=1}^{4}|\psi_j|^2=1 \tag{15}$$

This constraint, also called the normalization condition of the wave function, keeps the norm of $|\psi(t)\rangle$ at unit length.

At the beginning of the IGT, the agent does not have any preference among the four decks. Therefore, we assign an equal amplitude to each action in the first trial, as with the equal expectancies at the beginning of the CRL models:

$$|\psi(1)\rangle=\sum_{j=1}^{4}\frac{1}{2}|a_j\rangle \tag{16}$$

Probability amplitude updating. After the agent observes the outcome of the action selection, he updates his estimation of the goodness of each deck. The updating algorithm, also known as the amplitude-amplification algorithm, lies at the core of QRL. Amplitude amplification is a technique in quantum computing that generalizes the idea behind Grover's search algorithm[95] and gives rise to a family of quantum algorithms. Here, one specific version of amplitude amplification was used[6] and further modified to accommodate our task.

We have two unitary operators based on the current chosen action $|a\rangle$:

$$\hat{Q}_1=\hat{I}-(1-e^{i\phi_1})|a\rangle\langle a| \tag{17}$$

$$\hat{Q}_2=(1-e^{i\phi_2})|\psi(t)\rangle\langle\psi(t)|-\hat{I} \tag{18}$$

where $\langle a|$ in the dual space represents the complex conjugate (Hermitian conjugate) of $|a\rangle$ and $\hat{I}$ is the identity operator. The exponents ϕ_1 and ϕ_2 are the

learning factors that control the amount of amplification or attenuation of the amplitude in each trial. These matrices are the generalization of the oracle O in the Grover algorithm[4]. A unitary operator $\hat{U}$ means $\hat{U}^{\dagger}\hat{U}=\hat{U}\hat{U}^{\dagger}=\hat{I}$. Thus, these operators do not change the unit length of any normalized state vector.

The Grover operator is then the unitary transformation:

$$\hat{U}_{\mathrm{G}}=\hat{Q}_2\hat{Q}_1 \tag{19}$$

After the unitary transformation operates L times, the amplitude vector in the next trial becomes:

$$|\psi(t+1)\rangle=\hat{U}_{\mathrm{G}}^{L}\,|\psi(t)\rangle \tag{20}$$

The parameters ϕ_1, ϕ_2, and L are determined by the current utility $u(t)$. We have at least two approaches to deal with this. One is to fix ϕ_1 and ϕ_2 (usually equal to π) and determine L using the current utility[4]. The other is to fix $L=1$ and calculate ϕ_1 and ϕ_2 from the current utility[6]. We chose the latter because the former approach has the disadvantage that the amplitude could jump only discretely. However, the way we set ϕ_1 and ϕ_2 is different:

$$u(t)=\begin{cases}\lambda_{\mathrm{gain}}x(t)^{\alpha} & \text{if} \quad x(t)\geq 0\\ -\lambda_{\mathrm{loss}}|x(t)|^{\alpha} & \text{if} \quad x(t)<0\end{cases} \tag{21}$$

where $x(t)$ is the net outcome, α ($0<\alpha<1$) is the shape parameter, λ_{gain} ($0<\lambda_{\mathrm{gain}}<2$) and λ_{loss} ($0<\lambda_{\mathrm{loss}}<2$) are the reward-seeking and loss-aversion parameters. A scale for pay-offs is also used here. Then ϕ_1 and ϕ_2 are defined as:

$$\phi_1=\pi[u(t)\cos\pi\eta+b_1] \tag{22}$$

$$\phi_2=\pi[u(t)\sin\pi\eta+b_2] \tag{23}$$

If we consider (ϕ_1, ϕ_2) as the vector-valued function of $u(t)$, then (ϕ_1, ϕ_2) delineates a line in two-dimensional Euclidean space, passing through the point (b_1, b_2) with slope $\tan\pi\eta$ ($-1<\eta<1$). We refer to the free parameters η, b_1, and b_2 as the learning parameters.

The QSPP model. Inspired by the hybrid VPP models, we also implemented a hybrid QSPP model that combined the QSL model and perseverance (Supplementary Fig. 2d). The probability of choosing deck j in trial t is defined as:

$$\Pr[\delta_j(t)=1]=w\Pr_{\mathrm{QSL}}[\delta_j(t)=1]+(1-w)\Pr_{\mathrm{pers}}[\delta_j(t)=1] \tag{24}$$

where w ($0<w<1$) is the QRL weight. $\Pr_{\mathrm{QSL}}$ is the probability determined by the QSL amplitude and $\Pr_{\mathrm{pers}}$ is the classical probability using the SoftMax trial-independent rule:

$$\Pr_{\mathrm{pers}}[\delta_j(t)=1]=\frac{\exp\{P_j(t)\}}{\sum_{k=1}^{4}\exp\{P_k(t)\}} \tag{25}$$

Here, the perseverance strength $P_j(t)$ on deck j is defined in a similar way as above:

$$P_j(t)=\begin{cases}A_{\mathrm{pers}}P_j(t-1) & \text{if} \quad \delta_j(t-1)=0\\ A_{\mathrm{pers}}P_j(t-1)+\epsilon_{\mathrm{p}} & \text{if} \quad \delta_j(t-1)=1 \quad \text{and} \quad x(t-1)\geq 0\\ A_{\mathrm{pers}}P_j(t-1)+\epsilon_{\mathrm{n}} & \text{if} \quad \delta_j(t-1)=1 \quad \text{and} \quad x(t-1)<0\end{cases} \tag{26}$$

but here both ϵ_{p} and ϵ_{n} are not free parameters. They are now determined by e_{p} and e_{n} according to:

$$\epsilon_{\mathrm{p}}=\begin{cases}3^{e_{\mathrm{p}}}-1 & \text{if } e_{\mathrm{p}}\geq 0\\ -3^{-e_{\mathrm{p}}}+1 & \text{if } e_{\mathrm{p}}<0\end{cases} \tag{27}$$

$$\epsilon_{\mathrm{n}}=\begin{cases}3^{e_{\mathrm{n}}}-1 & \text{if } e_{\mathrm{n}}\geq 0\\ -3^{-e_{\mathrm{n}}}+1 & \text{if } e_{\mathrm{n}}<0\end{cases} \tag{28}$$

where e_{p} ($-5<e_{\mathrm{p}}<5$) and e_{n} ($-5<e_{\mathrm{n}}<5$) are designed to absorb the consistency parameter c into ϵ_{p} and ϵ_{n}.

One-term and two-term models. For convenience, the VPP and QSPP models are described as two-term models because they have both the reinforcement learning part and the perseverance part, while the others (EVL, PVL and QSL) are described as one-term models. The baseline, the EVL, the PVL, the QSL, the VPP and the QSPP models have 3, 4, 4, 6, 8 and 10 free parameters, respectively.

Statistical analyses. We used an alpha level of 0.05 for all behavioural statistical tests reported.

Maximum likelihood estimation. The parameters in each model were estimated by maximizing the LL of each subject's one-step-ahead predictions. The LL of subject i and model m is defined as:

$$\begin{aligned}\mathrm{LL}_i^m &= \log\Pr{}^m[\text{data for subject } i]\\ &= \sum_{t=1}^{T}\sum_{j=1}^{D}\delta_j(t)\log\Pr{}^m[\delta_j(t)=1|O_i(1),\cdots,O_i(t-1)]\end{aligned} \tag{29}$$

where T is the number of trials, D is the number of decks and $O_i(t)$ is the actual choice and consequent outcome in trial t. We used an optimization method called Bayesian Adaptive Direct Search that showed a comparable or even better performance than other common optimization methods for behavioural modelling[96].

Model comparison with the goodness-of-fit criterion. For the goodness-of-fit criterion, the corrected AICc[36] and the BIC[37] provide a direct assessment of the one-step-ahead predictions. The AICc avoids the disadvantage of the Akaike's Information Criterion[97] when facing a small sample size. The BIC more strongly penalizes the number of parameters. For each model i, the AICc_i is:

$$\mathrm{AICc}_i=-2\log L_i+2K_i+\frac{2K_i^2+2K_i}{T-K_i-1} \tag{30}$$

and the BIC_i is:

$$\mathrm{BIC}_i=-2\log L_i+K_i\log T \tag{31}$$

where T is the total number of trials, L_i is the maximum LL for model i and K_i is the number of free parameters in the model. A smaller AICc or BIC value represents a better fit to the data. Average values for each model are calculated in each group.

Bayesian model comparison with variational Bayes method. We performed the Bayesian model comparison using the variational Bayes algorithm[38]. In the Bayesian graph, by treating the model as a random variable, the parameters from a Dirichlet distribution over models was estimated. Given the parameters, the distribution of a multinomial variable was used to describe the probability that one specific model generated the data from one specific subject. We entered $-\mathrm{AICc}/2$ and $-\mathrm{BIC}/2$ as the model log evidence into the estimation iteration process. The expected likelihood measures how likely it is that one model will generate a randomly selected subject's behaviour, where the exceedance probability describes how one model is more likely than all other models.

Simulation method. Unlike the goodness-of-fit criterion, the simulation method[30] was designed to assess the accuracy of one model generating predictions for entire choice sequences according to model parameters rather than the subjects' choice history. The computer agent uses the parameters estimated from the subjects' behaviour to perform the IGT several times according to the corresponding model. Here, for each subject and each model, we used the best 10 combinations of estimated parameters to run 100 simulations for total trials, producing 1,000 performing sequences. The combination of parameters that generated the best simulation performance was chosen for the following analyses. We then defined two types of MSE for these simulations. For all the MSEs defined below, standard errors were calculated among all subjects.

First, we let $B_{i,j}^{0}$ denote the frequency of selecting deck j averaging over all T trials for subject i and $B_{i,j}^{m}$ denote the frequency of selecting deck j averaging over all T trials and 100 simulations for an agent using the parameters of model m estimated from subject i. We have the MSE of the PoC for each deck:

$$\mathrm{MSE}_{\mathrm{PoC}}^{m}=\frac{1}{ND}\sum_{i=1}^{N}\sum_{j=1}^{D}(B_{i,j}^{m}-B_{i,j}^{0})^2 \tag{32}$$

where N is the number of subjects and D is the number of decks.

Second, $P_{i,j,b}^{0}$ is the frequency of selecting deck j of subject i during block b; $P_{i,j,b}^{m}$ is the frequency of selecting deck j during block b of an agent using model m with estimated parameters from subject i. The MSE of EDS curve is defined as:

$$\mathrm{MSE}_{\mathrm{EDS}}^{m}=\frac{1}{NDB}\sum_{i=1}^{N}\sum_{j=1}^{D}\sum_{b=1}^{B}(P_{i,j,b}^{m}-P_{i,j,b}^{0})^2 \tag{33}$$

where B is the number of blocks (one block contains 20 trials).

fMRI data acquisition. Gradient echo-planar MRI data were acquired during the whole task procedure using a 3 Tesla Siemens Magnetom Trio scanner (Siemens

Medical Solutions) in the Anhui Provincial Hospital. A circularly polarized head coil was used. The fMRI images were collected using a T2*-weighted echo-planar imaging sequence (repetition time = 2,000 ms, echo time = 30 ms, field of view = 240 mm, matrix = 64 × 64, flip angle = 85°) with 33 axial slices (no gaps, voxel size: 3.75 × 3.75 × 3.70 mm^3) covering the whole brain. Before entering the scanner, all subjects were asked to not move their head during all scans. Three functional scan runs of 420 s occurred during the IGT. Between each scan run, there was an interval of about 1 min. High-resolution T1-weighted three-dimensional gradient-echo images were also obtained (repetition time = 1,900 ms; echo time = 2.26 ms; inversion time = 900 ms; 1-mm isotropic voxel; 250-mm field of view) for stereotaxic transformation.

fMRI data preprocessing. The imaging data were analysed with Analysis of Functional Neuroimages (AFNI-17.3.01)[98]. As in the previous study[56,99], each subject's raw data were corrected for temporal shifts between slices and for motion using the midmost sub-brick as the base, spatially smoothed with a Gaussian kernel (full width at half maximum = 8 mm) and temporally normalized (for each voxel, the signal of each sub-brick was divided by the temporally averaged signal). Then images were normalized to the Talairach coordinate.

Data from eight subjects in the control group and nine subjects in the smoking group were discarded because of relatively large head movement in the MRI scanner (more than 2 mm or 2°) or scanning technique issues. Therefore, data from 50 subjects in the control group and 34 in the smoking group were included in the following fMRI whole-brain analysis.

fMRI data analyses. We implemented seven GLMs to analyse the fMRI data. In GLM1, for the QSPP model, we included the regressor of the current quantum transition amplitude at the outcome period. GLM1 also included two additional regressors corresponding to the stimulus and the outcome display, six additional regressors for head motion, one additional regressor for the reaction time and one nuisance regressor for onsets of no-response trials. Standard GLM analysis was performed using a gamma haemodynamic response function model. In GLM2, we included the regressor of quantum distance at the outcome period and also other additional or nuisance regressors. In GLM3, we included the regressor of generalized quantum distance at the outcome period and also other additional or nuisance regressors.

In GLM4, we included the outcome, the loss (penalty), uncertainty, the interaction of uncertainty and loss (penalty) and the current choice probability as the regressors at the outcome period and also other additional or nuisance regressors. The differences of each regressor between two models in the control group (control QSPP − control VPPDecayTIC) and the smoking group (smoking QSPP − smoking VPPDecayTIC) were calculated. In GLM5, we included the outcome, the gain (reward), the uncertainty, the interaction of uncertainty and gain (reward), the current choice probability and also other additional or nuisance regressors. In GLM6, we included the generalized quantum distance, the uncertainty, the interaction of uncertainty and generalized quantum distance and also other additional or nuisance regressors. In GLM7, we included the reward prediction error and the current action value provided by the VPPDecayTIC model as the regressors at the outcome period and also other additional or nuisance regressors. The results of GLM3–5 are reported in the main text while the rest are in the Supplementary Results. All reported voxels survived the family-wise error correction at a cluster level threshold of $P < 0.05$, with $P_{uncorrected} < 0.001$ (two-sided threshold cluster size = 65).

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All data are available from the corresponding author on reasonable request.

Code availability

All code used to generate the results central to the main claims in this study is available from the corresponding author on reasonable request.

Received: 19 December 2018; Accepted: 2 December 2019;
Published online: 20 January 2020

References

1. Sutton, R. S. & Barto, A. G. *Reinforcement Learning: An Introduction*, Vol. 1 (MIT Press, 1998).
2. Niv, Y. Reinforcement learning in the brain. *J. Math. Psychol.* **53**, 139–154 (2009).
3. Biamonte, J. et al. Quantum machine learning. *Nature* **549**, 195–202 (2017).
4. Dong, D., Chen, C., Li, H. & Tarn, T.-J. Quantum reinforcement learning. *IEEE Trans. Syst. Man Cybern. Pt B* **38**, 1207–1220 (2008).
5. Dong, D., Chen, C., Chu, J. & Tarn, T.-J. Robust quantum-inspired reinforcement learning for robot navigation. *IEEE/ASME Trans. Mechatron.* **17**, 86–97 (2012).
6. Fakhari, P., Rajagopal, K., Balakrishnan, S. N. & Busemeyer, J. R. Quantum inspired reinforcement learning in changing environment. *New Math. Nat. Comput.* **9**, 273–294 (2013).
7. Wittek, P. *Quantum Machine Learning: What Quantum Computing Means to Data Mining* (Academic Press, 2014).
8. Dunjko, V., Taylor, J. M. & Briegel, H. J. Quantum-enhanced machine learning. *Phys. Rev. Lett.* **117**, 130501 (2016).
9. Manousakis, E. Quantum formalism to describe binocular rivalry. *Biosystems* **98**, 57–66 (2009).
10. Busemeyer, J. R. & Bruza, P. D. *Quantum Models of Cognition and Decision* (Cambridge Univ. Press, 2012).
11. Busemeyer, J. R., Wang, Z. & Shiffrin, R. M. Bayesian model comparison favors quantum over standard decision theory account of dynamic inconsistency. *Decision* **2**, 1–12 (2015).
12. Kvam, P. D., Pleskac, T. J., Yu, S. & Busemeyer, J. R. Interference effects of choice on confidence: quantum characteristics of evidence accumulation. *Proc. Natl Acad. Sci. USA* **112**, 10645–10650 (2015).
13. Ashtiani, M. & Azgomi, M. A. A survey of quantum-like approaches to decision making and cognition. *Math. Soc. Sci.* **75**, 49–80 (2015).
14. Yukalov, V. I. & Sornette, D. Quantum probability and quantum decision-making. *Phil. Trans. R. Soc. A* **374**, 20150100 (2016).
15. de Barros, J. A. & Oas, G. in *The Palgrave Handbook of Quantum Models in Social Science* (eds Haven, E. & Khrennikov, A.) 195–228 (Springer, 2017).
16. Takahashi, T. Can quantum approaches benefit biology of decision making? *Prog. Biophys. Mol. Biol.* **130**, 99–102 (2017).
17. Gold, J. I. & Shadlen, M. N. The neural basis of decision making. *Annu. Rev. Neurosci.* **30**, 535–574 (2007).
18. Sanfey, A. G., Loewenstein, G., McClure, S. M. & Cohen, J. D. Neuroeconomics: cross-currents in research on decision-making. *Trends Cogn. Sci.* **10**, 108–116 (2006).
19. Glimcher, P. W. Indeterminacy in brain and behavior. *Annu. Rev. Psychol.* **56**, 25–56 (2005).
20. Glimcher, P. W. & Fehr, E. *Neuroeconomics: Decision Making and the Brain* (Academic Press, 2013).
21. Lee, D., Seo, H. & Jung, M. W. Neural basis of reinforcement learning and decision making. *Annu. Rev. Neurosci.* **35**, 287–308 (2012).
22. Daw, N. D. & Tobler, P. N. in *Neuroeconomics* 2nd edn (eds Glimcher, P. W. & Fehr, E.) 283–298 (Academic Press, 2014).
23. Kornmeier, J., Friedel, E., Wittmann, M. & Atmanspacher, H. EEG correlates of cognitive time scales in the Necker-Zeno model for bistable perception. *Conscious. Cogn.* **53**, 136–150 (2017).
24. Bechara, A., Damasio, A. R., Damasio, H. & Anderson, S. W. Insensitivity to future consequences following damage to human prefrontal cortex. *Cognition* **50**, 7–15 (1994).
25. Ahn, W. Y., Dai, J., Vassileva, J., Busemeyer, J. R. & Stout, J. C. in *Progress in Brain Research* Vol. 224 (eds Ekhtiari, H. & Paulus, M.) 53–65 (Elsevier, 2016).
26. Buelow, M. T. & Suhr, J. A. Risky decision making in smoking and nonsmoking college students: examination of Iowa Gambling Task performance by deck type selections. *Appl. Neuropsychol. Child* **3**, 38–44 (2014).
27. Wei, Z. et al. Chronic nicotine exposure impairs uncertainty modulation on reinforcement learning in anterior cingulate cortex and serotonin system. *NeuroImage* **169**, 323–333 (2018).
28. Steingroever, H. et al. Data from 617 healthy participants performing the Iowa gambling task: a "many labs" collaboration. *J. Open Psychol. Data* **3**, 340–353 (2015).
29. Rangel, A., Camerer, C. & Montague, P. R. A framework for studying the neurobiology of value-based decision making. *Nat. Rev. Neurosci.* **9**, 545–556 (2008).
30. Ahn, W.-Y., Busemeyer, J. R., Wagenmakers, E.-J. & Stout, J. C. Comparison of decision learning models using the generalization criterion method. *Cogn. Sci.* **32**, 1376–1402 (2008).
31. Worthy, D. A., Pang, B. & Byrne, K. A. Decomposing the roles of perseveration and expected value representation in models of the Iowa gambling task. *Front. Psychol.* **4**, 640 (2013).
32. Ahn, W. Y. et al. Decision-making in stimulant and opiate addicts in protracted abstinence: evidence from computational modeling with pure users. *Front. Psychol.* **5**, 849 (2014).
33. Worthy, D. A. & Maddox, W. T. Age-based differences in strategy use in choice tasks. *Front. Neurosci.* **5**, 145 (2012).
34. Ahn, W.-Y., Krawitz, A., Kim, W., Busemeyer, J. R. & Brown, J. W. A model-based fMRI analysis with hierarchical Bayesian parameter estimation. *Decision* **1**, 8–23 (2013).
35. Byrne, K. A., Norris, D. D. & Worthy, D. A. Dopamine, depressive symptoms, and decision-making: the relationship between spontaneous eye blink rate and depressive symptoms predicts Iowa Gambling Task performance. *Cogn. Affect. Behav. Neurosci.* **16**, 23–36 (2016).
36. Cavanaugh, J. E. Unifying the derivations for the Akaike and corrected Akaike information criteria. *Stat. Probab. Lett.* **33**, 201–208 (1997).

37. Schwarz, G. Estimating the dimension of a model. *Ann. Stat.* **6**, 461–464 (1978).
38. Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J. & Friston, K. J. Bayesian model selection for group studies. *NeuroImage* **46**, 1004–1017 (2009).
39. Dajka, J., Łuczka, J. & Hänggi, P. Distance between quantum states in the presence of initial qubit-environment correlations: a comparative study. *Phys. Rev. A* **84**, 032120 (2011).
40. O'Doherty, J. P., Hampton, A. & Kim, H. Model-based fMRI and its application to reward learning and decision making. *Ann. N. Y. Acad. Sci.* **1104**, 35–53 (2007).
41. Ma, W. J. & Jazayeri, M. Neural coding of uncertainty and probability. *Annu. Rev. Neurosci.* **37**, 205–220 (2014).
42. Bach, D. R., Hulme, O., Penny, W. D. & Dolan, R. J. The known unknowns: neural representation of second-order uncertainty, and ambiguity. *J. Neurosci.* **31**, 4811–4820 (2011).
43. Payzan-LeNestour, E., Dunne, S., Bossaerts, P. & O'Doherty, J. P. The neural representation of unexpected uncertainty during value-based decision making. *Neuron* **79**, 191–201 (2013).
44. Behrens, T. E. J., Woolrich, M. W., Walton, M. E. & Rushworth, M. F. S. Learning the value of information in an uncertain world. *Nat. Neurosci.* **10**, 1214–1221 (2007).
45. Yu, A. J. & Dayan, P. Uncertainty, neuromodulation, and attention. *Neuron* **46**, 681–692 (2005).
46. Singh, V. A potential role of reward and punishment in the facilitation of the emotion-cognition dichotomy in the Iowa Gambling Task. *Front. Psychol.* **4**, 944 (2013).
47. Yechiam, E. & Ert, E. Evaluating the reliance on past choices in adaptive learning models. *J. Math. Psychol.* **51**, 75–84 (2007).
48. Chuang, I. L., Gershenfeld, N. & Kubinec, M. Experimental implementation of fast quantum searching. *Phys. Rev. Lett.* **80**, 3408 (1998).
49. Dunjko, V., Taylor, J. M. & Briegel, H. J. Advances in quantum reinforcement learning. In *Proc. 2017 IEEE International Conference on Systems, Man, and Cybern*etics 282–287 (IEEE, 2017).
50. Nielsen, M. A. & Chuang, I. L. *Quantum Computation and Quantum Information* (Cambridge Univ. Press, 2010).
51. Yearsley, J. M. Advanced tools and concepts for quantum cognition: a tutorial. *J. Math. Psychol.* **78**, 24–39 (2017).
52. Crawford, D., Levit, A., Ghadermarzy, N., Oberoi, J. S. & Ronagh, P. Reinforcement learning using quantum Boltzmann machines. *Quantum Info. Comput.* **18**, 51–74 (2018).
53. Krain, A. L., Wilson, A. M., Arbuckle, R., Castellanos, F. X. & Milham, M. P. Distinct neural mechanisms of risk and ambiguity: a meta-analysis of decision-making. *NeuroImage* **32**, 477–484 (2006).
54. Hsu, M., Bhatt, M., Adolphs, R., Tranel, D. & Camerer, C. F. Neural systems responding to degrees of uncertainty in human decision-making. *Science* **310**, 1680–1683 (2005).
55. Litt, A., Plassmann, H., Shiv, B. & Rangel, A. Dissociating valuation and saliency signals during decision-making. *Cereb. Cortex* **21**, 95–102 (2010).
56. Wang, Y. et al. Neural substrates of updating the prediction through prediction error during decision making. *NeuroImage* **157**, 1–12 (2017).
57. Vickery, T. J. & Jiang, Y. V. Inferior parietal lobule supports decision making under uncertainty in humans. *Cereb. Cortex* **19**, 916–925 (2008).
58. Xue, G., Lu, Z., Levin, I. P. & Bechara, A. The impact of prior risk experiences on subsequent risky decision-making: the role of the insula. *NeuroImage* **50**, 709–716 (2010).
59. Haggard, P. Human volition: towards a neuroscience of will. *Nat. Rev. Neurosci.* **9**, 934–946 (2008).
60. Nachev, P., Kennard, C. & Husain, M. Functional role of the supplementary and pre-supplementary motor areas. *Nat. Rev. Neurosci.* **9**, 856–869 (2008).
61. Tanji, J. & Kurata, K. Contrasting neuronal activity in supplementary and precentral motor cortex of monkeys. I. Responses to instructions determining motor responses to forthcoming signals of different modalities. *J. Neurophysiol.* **53**, 129–141 (1985).
62. Okano, K. & Tanji, J. Neuronal activities in the primate motor fields of the agranular frontal cortex preceding visually triggered and self-paced movement. *Exp. Brain Res.* **66**, 155–166 (1987).
63. Rushworth, M. F. S. & Behrens, T. E. J. Choice, uncertainty and value in prefrontal and cingulate cortex. *Nat. Neurosci.* **11**, 389–397 (2008).
64. Sul, J. H., Kim, H., Huh, N., Lee, D. & Jung, M. W. Distinct roles of rodent orbitofrontal and medial prefrontal cortex in decision making. *Neuron* **66**, 449–460 (2010).
65. Kepecs, A., Uchida, N., Zariwala, H. A. & Mainen, Z. F. Neural correlates, computation and behavioural impact of decision confidence. *Nature* **455**, 227–231 (2008).
66. O'Neill, M. & Schultz, W. Coding of reward risk by orbitofrontal neurons is mostly distinct from coding of reward value. *Neuron* **68**, 789–800 (2010).
67. Studer, B., Cen, D. & Walsh, V. The angular gyrus and visuospatial attention in decision-making under risk. *NeuroImage* **103**, 75–80 (2014).
68. Tversky, A. & Kahneman, D. Advances in prospect theory: cumulative representation of uncertainty. *J. Risk Uncertain.* **5**, 297–323 (1992).
69. De Barros, J. A. & Suppes, P. Quantum mechanics, interference, and the brain. *J. Math. Psychol.* **53**, 306–313 (2009).
70. Lambert, N. et al. Quantum biology. *Nat. Phys.* **9**, 10–18 (2013).
71. Busemeyer, J. R., Pothos, E. M., Franco, R. & Trueblood, J. S. A quantum theoretical explanation for probability judgment errors. *Psychol. Rev.* **118**, 193–218 (2011).
72. beim Graben, P. & Atmanspacher, H. Complementarity in classical dynamical systems. *Found. Phys.* **36**, 291–306 (2006).
73. beim Graben, P., Filk, T. & Atmanspacher, H. Epistemic entanglement due to non-generating partitions of classical dynamical systems. *Int. J. Theor. Phys.* **52**, 723–734 (2013).
74. Ivakhnenko, O. V., Shevchenko, S. N. & Nori, F. Simulating quantum dynamical phenomena using classical oscillators: Landau-Zener-Stückelberg-Majorana interferometry, latching modulation, and motional averaging. *Sci. Rep.* **8**, 12218 (2018).
75. Bliokh, K. Y., Bekshaev, A. Y., Kofman, A. G. & Nori, F. Photon trajectories, anomalous velocities and weak measurements: a classical interpretation. *New J. Phys.* **15**, 073022 (2013).
76. Carleo, G. & Troyer, M. Solving the quantum many-body problem with artificial neural networks. *Science* **355**, 602–606 (2017).
77. Busemeyer, J. R., Fakhari, P. & Kvam, P. Neural implementation of operations used in quantum cognition. *Prog. Biophys. Mol. Biol.* **130**, 53–60 (2017).
78. Phelps, E. A., Lempert, K. M. & Sokol-Hessner, P. Emotion and decision making: multiple modulatory neural circuits. *Annu. Rev. Neurosci.* **37**, 263–287 (2014).
79. Hu, H. Reward and aversion. *Annu. Rev. Neurosci.* **39**, 297–324 (2016).
80. Chen, C., Takahashi, T., Nakagawa, S., Inoue, T. & Kusumi, I. Reinforcement learning in depression: a review of computational research. *Neurosci. Biobehav. Rev.* **55**, 247–267 (2015).
81. Sanfey, A. G. Social decision-making: insights from game theory and neuroscience. *Science* **318**, 598–602 (2007).
82. Roskies, A. L. How does neuroscience affect our conception of volition? *Annu. Rev. Neurosci.* **33**, 109–130 (2010).
83. Schack, R., Brun, T. A. & Caves, C. M. Quantum Bayes rule. *Phys. Rev. A* **64**, 014305 (2001).
84. Kouda, N., Matsui, N., Nishimura, H. & Peper, F. Qubit neural network and its learning efficiency. *Neural Comput. Appl.* **14**, 114–121 (2005).
85. Piotrowski, E. W. & Sladkowski, J. The next stage: quantum game theory. in *Mathematical Physics Research at the Cutting Edge* (ed. Benton, C. V.) 247–268 (*Nova Science Publishers*, 2004).
86. Ahn, W.-Y., Krawitz, A., Kim, W., Busemeyer, J. R. & Brown, J. W. A model-based fMRI analysis with hierarchical Bayesian parameter estimation. *J. Neurosci. Psychol. Econ.* **4**, 95–110 (2011).
87. He, Q. et al. Altered dynamics between neural systems sub-serving decisions for unhealthy food. *Front. Neurosci.* **8**, 350 (2014).
88. Brevers, D., Noël, X., He, Q., Melrose, J. A. & Bechara, A. Increased ventral-striatal activity during monetary decision making is a marker of problem poker gambling severity. *Addict. Biol.* **21**, 688–699 (2016).
89. Yechiam, E. & Busemeyer, J. R. Comparison of basic assumptions embedded in learning models for experience-based decision making. *Psychon. Bull. Rev.* **12**, 387–402 (2005).
90. Busemeyer, J. R. & Stout, J. C. A contribution of cognitive decision models to clinical assessment: decomposing performance on the Bechara gambling task. *Psychol. Assess.* **14**, 253–262 (2002).
91. Erev, I. & Barron, G. On adaptation, maximization, and reinforcement learning among cognitive strategies. *Psychol. Rev.* **112**, 912–931 (2005).
92. Ahn, W.-Y., Haines, N. & Zhang, L. Revealing neurocomputational mechanisms of reinforcement learning and decision-making with the hBayesDM package. *Comput. Psychiatr.* **1**, 24–57 (2017).
93. Wagner, A. R. & Rescorla, R. A. in *Inhibition and Learning* (eds Boakes, R. A. & Halliday, M. S.) 301–336 (1972).
94. Erev, I. & Roth, A. E. Predicting how people play games: reinforcement learning in experimental games with unique, mixed strategy equilibria. *Am. Econ. Rev.* **88**, 848–881 (1998).
95. Grover, L. K. A fast quantum mechanical algorithm for database search. In *Proc. 28th Annual ACM Symposium on Theory of Computing* 212–219 (ACM, 1996).
96. Acerbi, L. & Ji, W. Practical Bayesian optimization for model fitting with Bayesian adaptive direct search. *Adv. Neural Inf. Proc. Syst.* **30**, 1836–1846 (2017).
97. Akaike, H. A new look at the statistical model identification. *IEEE Trans. Automat. Contr.* **19**, 716–723 (1974).
98. Cox, R. W. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Comput. Biomed. Res.* **29**, 162–173 (1996).
99. Li, N. et al. Resting-state functional connectivity predicts impulsivity in economic decision-making. *J. Neurosci.* **33**, 4886–4895 (2013).

Acknowledgements

We thank Y. Yang, R. Zha, J. Besumeyer and N. Ma for their inspirational comments. We thank L. Acerbi, G. R. Yang, C. Gneiting, A. Miranowicz, X. Li, Z. Jin and X. Li for their helpful suggestions. This work was supported by grants from the National Key Basic Research Programme (grant nos. 2016YFA0400900 and 2018YFC0831101), the National Natural Science Foundation of China (grant nos. 31471071, 31771221, 61773360, 71671115, 71874170 and 71942003), the Fundamental Research Funds for the Central Universities of China, the MURI Center for Dynamic Magneto-Optics via the Air Force Office of Scientific Research (AFOSR; grant no. FA9550-14-1-0040), the Army Research Office (ARO; grant no. W911NF-18-1-0358), the Asian Office of Aerospace Research and Development (AOARD; grant no. FA2386-18-1-4045), the Japan Science and Technology Agency (JST; via the Q-LEAP programme and CREST grant no. JPMJCR1676), the Japan Society for the Promotion of Science (JSPS; JSPS–RFBR grant no. 17-52-50023 and JSPS–FWO grant no. VS.059.18N), the RIKEN–AIST Challenge Research Fund, the Templeton Foundation, the Foundational Questions Institute (FQXi) and the NTT PHI Laboratory, the Australian Research Council's Discovery Projects funding scheme under Project DP190101566, the Alexander von Humboldt Foundation and the US Office of Naval Research. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. We thank the Bioinformatics Centre of the University of Science and Technology of China, School of Life Science for providing supercomputing resources for this project.

Author contributions

L.J.-A., Y.P. and X.Z. conceived the study. Y.L. and Z.W. provided the devices and collected the data. L.J.-A. built the models. L.J.-A. and Z.W. analysed the data. All authors participated in discussions. L.J.-A., D.D., Y.P., F.N. and X.Z. wrote the paper. X.Z. supervised the project and acquired funding.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41562-019-0804-2.

Correspondence and requests for materials should be addressed to X.Z.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Primary Handling Editor: Stavroula Kousta.