nature human behaviour

SUPPLEMENTARY INFORMATION

https://doi.org/10.1038/s41562-019-0804-2

In the format provided by the authors and unedited.

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]*

[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

Supplementary Information
Quantum Reinforcement Learning during Human Decision Making

Contents

Supplementary Methods

Further explanations of the quantum operator $\hat{U}_{\mathbf{G}}$

Math Preparation

In order to illustrate the effect of the operator $\hat{U}_{\mathrm{G}}$ in the QSL model, we rewrite the superposition state $|\psi\rangle$ in trial t in this form (t is omitted for simplicity):

$$\begin{aligned} |\psi\rangle &= \sum_{k=1}^{4} \psi_k \, |a_k\rangle \\ &= \psi_a \, |a\rangle + \psi_{a_\perp} \, |a_\perp\rangle \ , \end{aligned} \tag{1}$$

where $|a\rangle$ is the chosen action, and $|a_\perp\rangle$ is a vector orthogonal to $|a\rangle$ satisfying

$$|a_\perp\rangle = \sum_{a_k \neq a} \frac{\psi_k}{\psi_{a_\perp}} \, |a_k\rangle \ , \tag{2}$$

where

$$\psi_{a_\perp} = \sqrt{\sum_{a_k \neq a} |\psi_k|^2} = \sqrt{1 - |\psi_a|^2} \, . \tag{3}$$

This means that $|\psi\rangle$ can be taken as a vector in the two dimensional space spanned by $|a\rangle$ and $|a_\perp\rangle$.

Geometric explanation for fixed learning factors

We first consider a simple example, where $\phi_1 = \phi_2 = \pi$:

$$\hat{Q}_1 = \hat{I} - 2\,|a\rangle\,\langle a|\ , \tag{4}$$

$$\hat{Q}_2 = 2\,|\psi\rangle\,\langle\psi| - \hat{I}\ . \tag{5}$$

It is clear that

$$\hat{Q}_1\,|a\rangle = (\hat{I} - 2\,|a\rangle\,\langle a|)\,|a\rangle = -\,|a\rangle\ , \tag{6}$$

$$\hat{Q}_1\,|a_\perp\rangle = (\hat{I} - 2\,|a\rangle\,\langle a|)\,|a_\perp\rangle = |a_\perp\rangle\ . \tag{7}$$

Here, $\hat{Q}_1$ flips the sign of the amplitude of action $|a\rangle$ but keeps the amplitude of any action orthogonal to $|a\rangle$. Hence $\hat{Q}_1$ will reflect any vector about the hyperplane orthogonal to $|a\rangle$. Similarly, $\hat{Q}_2$ keeps the sign of the amplitude of action $|\psi\rangle$ but flips the amplitude of any action orthogonal to $|\psi\rangle$.

Let $|\langle\psi|a\rangle| = \sin\bar{\theta}$ (having a period of 2π). $\hat{Q}_1$ first flips $|\psi\rangle$ into $|\psi'\rangle = \hat{Q}_1\,|\psi\rangle$ (Figure S3), and $\hat{Q}_2$ then flips $|\psi'\rangle$ into $|\psi''\rangle = \hat{Q}_2\,|\psi'\rangle$. The pure effect of $\hat{U}_{\mathrm{G}}$ is rotating $|\psi\rangle$ by $2\bar{\theta}$. This example is slightly different from the original one[1], where $\hat{Q}_2$ is constructed from another vector with different properties.

Geometric explanation for the general case

We now go to the full formula of Grover iteration, where the learning factors are flexible:

$$\hat{Q}_1 = \hat{I} - (1 - e^{i\phi_1})\,|a\rangle\,\langle a|\ , \tag{8}$$

$$\hat{Q}_2 = (1 - e^{i\phi_2}) \left|\psi\right\rangle \left\langle\psi\right| - \hat{I} \ . \tag{9}$$

Similarly, we have

$$\hat{Q}_1 \left|a\right\rangle = (\hat{I} - (1 - e^{i\phi_1}) \left|a\right\rangle \left\langle a\right|) \left|a\right\rangle = e^{i\phi_1} \left|a\right\rangle \ , \tag{10}$$

$$\hat{Q}_1 \left|a_\perp\right\rangle = (\hat{I} - (1 - e^{i\phi_1}) \left|a\right\rangle \left\langle a\right|) \left|a_\perp\right\rangle = \left|a_\perp\right\rangle \ , \tag{11}$$

so that,

$$\begin{aligned} \left|\psi'\right\rangle &= \hat{Q}_1 \left|\psi\right\rangle \\ &= \left[\hat{I} - (1 - e^{i\phi_1}) \left|a\right\rangle \left\langle a\right|\right] \left|\psi\right\rangle \\ &= e^{i\phi_1} \psi_a \left|a\right\rangle + \psi_{a_\perp} \left|a_\perp\right\rangle \ , \end{aligned} \tag{12}$$

where $\hat{Q}_1$ acts as a phase gate (conditional phase shift operation) in quantum computation,

$$U_{\text{phase}} = \begin{bmatrix} e^{i\phi} & 0 \\ 0 & 1 \end{bmatrix} \ . \tag{13}$$

To visualize the transformation geometrically, we introduce the Bloch sphere representation[2], which provides a useful means to visualize a single qubit (like our $\left|\psi\right\rangle$ in the two dimensional Hilbert space). We have

$$\begin{aligned} \left|\psi\right\rangle &= \psi_a \left|a\right\rangle + \psi_{a_\perp} \left|a_\perp\right\rangle \\ &= e^{i\gamma}(\cos\frac{\theta}{2} \left|a\right\rangle + e^{i\varphi} \sin\frac{\theta}{2} \left|a_\perp\right\rangle) \\ &\simeq \cos\frac{\theta}{2} \left|a\right\rangle + e^{i\varphi} \sin\frac{\theta}{2} \left|a_\perp\right\rangle \ , \end{aligned} \tag{14}$$

where the factor $e^{i\gamma}$ can be ignored, because a global phase has no observable effects. The parameter polar angle θ (having a period of π) and azimuthal angle φ (having a period of 2π) define a point on the Bloch sphere (Figure 2c in the main text). A Bloch sphere is a two-dimensional manifold embedded in the three-dimensional Euclidean space, with antipodal points corresponding to a pair of orthogonal vectors. The vectors or states $|a\rangle$ and $|a_\perp\rangle$ are the north and south poles (zenith direction) and the angle between them is π, different from $\pi/2$ in Figure S3. Here, θ is conceptually equivalent to $2\bar{\theta}$ in Figure S3.

The effect that $\hat{Q}_1$ adds ϕ_1 to the phase of the amplitude of $|a\rangle$ can be seen as subtracting ϕ_1 from the phase φ of the amplitude of $|a_\perp\rangle$ if throwing away a global phase, and thus can be shown as the clockwise rotation around the z-axis $\hat{z}$ by ϕ_1 on the blue circle on the Bloch sphere, without changing the value of θ, rotating $|\psi\rangle$ into $|\psi'\rangle$.

We then transform the basis $\{|a\rangle\,,\ |a_\perp\rangle\}$ into the basis $\{|\psi\rangle\,,\ |\psi_\perp\rangle\}$, and $|\psi\rangle$ becomes the new z-axis $\hat{z}'$. Similarly, $|\psi'\rangle$ has the two spherical coordinates parameters θ' and φ'. $\hat{Q}_2$ also subtracts ϕ_2 from the phase φ' of the amplitude of $|\psi_\perp\rangle$, which clockwise rotates around the new z-axis $\hat{z}'$ on the purple circle, rotating $|\psi'\rangle$ into $|\psi''\rangle$.

Taken together, the effect of $\hat{U}_{\mathrm{G}}$ is a two-step rotation, which finally changes the angle θ in the basis $\{|a\rangle\,,\ |a_\perp\rangle\}$. The composition of two rotations is still a rotation. To determine the rotation angle and the rotation axis, we use the notation of three-dimensional rotations[3]. Assuming that the first rotation is $\beta\hat{m}$ (rotating around $\hat{m}$ by β) and the second is $\alpha\hat{l}$, then the composition rotation is

$\gamma\hat{n}$. We have

$$\begin{aligned}\cos\frac{\gamma}{2} &= \cos\frac{\alpha}{2}\cos\frac{\beta}{2} - \sin\frac{\alpha}{2}\sin\frac{\beta}{2}\hat{l}\cdot\hat{m} \\ \sin\frac{\gamma}{2}\hat{n} &= \sin\frac{\alpha}{2}\cos\frac{\beta}{2}\hat{l} + \cos\frac{\alpha}{2}\sin\frac{\beta}{2}\hat{m} + \sin\frac{\alpha}{2}\sin\frac{\beta}{2}\hat{l}\times\hat{m}\ .\end{aligned} \tag{15}$$

Taking $\alpha = -\phi_2$, $\beta = -\phi_1$, $\hat{l}$ and $\hat{m}$ to be the direction vector of $|\psi\rangle$ and $|a\rangle$, we obtain the composition results immediately. Naturally, a generalization of Grover iteration could be an arbitrary parametric rotation on the Bloch sphere.

Visualization of the general case

We can also compute the effect of $\hat{U}_{\mathrm{G}}$ analytically:

$$\begin{aligned}\hat{Q}_1|\psi\rangle &= \Big[\hat{I} - (1-e^{i\phi_1})|a\rangle\langle a|\Big]|\psi\rangle \\ &= e^{i\phi_1}\psi_a|a\rangle + \psi_{a_\perp}|a_\perp\rangle\end{aligned} \tag{16}$$

$$\begin{aligned}\hat{Q}_2\hat{Q}_1|\psi\rangle &= (1-e^{i\phi_2})\Big[\psi_a|a\rangle + \psi_{a_\perp}|a_\perp\rangle\Big]\Big[\psi_a^*\langle a| + \psi_{a_\perp}^*\langle a_\perp|\Big]\hat{Q}_1|\psi\rangle - \hat{Q}_1|\psi\rangle \\ &= (f-e^{i\phi_1})\psi_a|a\rangle + (f-1)\psi_{a_\perp}|a_\perp\rangle\ ,\end{aligned} \tag{17}$$

where

$$\begin{aligned}f &= (1-e^{i\phi_2})(e^{i\phi_1}|\psi_a|^2 + |\psi_{a_\perp}|^2) \\ &= (1-e^{i\phi_2})(e^{i\phi_1}|\psi_a|^2 + 1 - |\psi_a|^2) \\ &= (1-e^{i\phi_2})\Big[1-(1-e^{i\phi_1})\Big]|\psi_a|^2\ .\end{aligned} \tag{18}$$

The effect of $\hat{U}_{\mathrm{G}}$ on $|\psi\rangle$ is to change the amplitude of $|a\rangle$ and $|a_{\perp}\rangle$. The ratio of the amplitude of action $|a\rangle$ after and before $\hat{U}_{\mathrm{G}}$ can be written as:

$$\begin{aligned} R &= f - e^{i\phi_1} \\ &= (1 - e^{i\phi_1} - e^{i\phi_2}) - (1 - e^{i\phi_1})(1 - e^{i\phi_2})|\psi_a|^2 \ . \end{aligned} \tag{19}$$

Let $p = |\psi_a|^2$ be the probability of action $|a\rangle$ before learning. Then the ratio of the probability after (p_{new}) and before (p) learning is:

$$\begin{aligned} |R|^2 =& \frac{p_{\text{new}}}{p} \\ =& \Big|(1 - e^{i\phi_1} - e^{i\phi_2}) - (1 - e^{i\phi_1})(1 - e^{i\phi_2})p\Big|^2 \\ =& 3 + 2\Big[\cos(\phi_1 - \phi_2) - \cos\phi_1 - \cos\phi_2\Big] \\ & - \Big[6 + 4\cos(\phi_1 - \phi_2) + 2\cos(\phi_1 + \phi_2) - 6\cos\phi_1 - 6\cos\phi_2\Big]p \\ & + 4(1 - \cos\phi_1)(1 - \cos\phi_2)p^2 \ , \end{aligned} \tag{20}$$

which is symmetric about $\phi_1 = \phi_2$ and $\phi_1 = -\phi_2$, and depends only on the learning factors ϕ_1, ϕ_2, and the current probability p. For convenience, we use R^2 to refer to $|R|^2$ later.

We can plot p_{new} or $LR = \log R^2$ as a function of ϕ_1, ϕ_2 and p (Figure S4a-b). A LR larger than zero means reinforcement (red) and smaller than zero means penalization (blue) (Figure S4b). The mapping from $u(t)$ to (ϕ_1, ϕ_2) then delineates a line with slope $\tan\pi\eta$ $(-1 < \eta < 1)$ moving from the starting point (b_1, b_2), the intersection point of the cyan (when the outcome is smaller than zero) and orange lines (when the outcome is larger than zero). The outcome determines one point on this line and reinforces or penalizes the action based on the LR value (Figure S4c-d). The previous work chose (η, b_1, b_2) manually[4]; in contrast, in our models, we left these parameters free

and optimized them to fit the data.

Quantum transition amplitudes and probabilities

In each trial t, since the Grover operation can be described in the two-dimensional space spanned by $|a\rangle$ and $|a_\perp\rangle$, all the operators have a two-dimensional representation in this basis. The representation of $|\psi\rangle$ is

$$\vec{\psi} = \begin{bmatrix} \psi_a \\ \\ \psi_{a_\perp} \end{bmatrix} . \tag{21}$$

And the representations of $\hat{Q}_1$, $\hat{Q}_2$ and $\hat{U}_{\rm G}$ are

$$Q_1 = \begin{bmatrix} e^{i\phi_1} & 0 \\ \\ 0 & 1 \end{bmatrix} , \tag{22}$$

$$Q_2 = \begin{bmatrix} (1-e^{i\phi_2})\psi_a\psi_a^* - 1 & (1-e^{i\phi_2})\psi_a\psi_{a_\perp}^* \\ \\ (1-e^{i\phi_2})\psi_{a_\perp}\psi_a^* & (1-e^{i\phi_2})\psi_{a_\perp}\psi_{a_\perp}^* - 1 \end{bmatrix} , \tag{23}$$

$$U_{\rm G} = Q_2Q_1 = \begin{bmatrix} e^{i\phi_1}[(1-e^{i\phi_2})\psi_a\psi_a^* - 1] & (1-e^{i\phi_2})\psi_a\psi_{a_\perp}^* \\ \\ e^{i\phi_1}(1-e^{i\phi_2})\psi_{a_\perp}\psi_a^* & (1-e^{i\phi_2})\psi_{a_\perp}\psi_{a_\perp}^* - 1 \end{bmatrix} . \tag{24}$$

From the theory of the special unitary group, it can be shown that the general expression of any 2×2 unitary matrix U is:

$$U = \begin{bmatrix} a & b \\ \\ -e^{i\phi}b^* & e^{i\phi}a^* \end{bmatrix} , \ |a|^2 + |b|^2 = 1 , \tag{25}$$

which depends on four free real parameters (the phase of a, the phase of b, the angle ϕ and the magnitude between a and b). However, if U is timed by $e^{i\bar{\phi}}$ for some $\bar{\phi}$, its effect does not change because a global phase shift is not observable.

In our case, a and b depict the effect of $\hat{U}_{\mathrm{G}}$ on $\vec{\psi}$. We have

$$\begin{aligned} |a|^2 &= \left|e^{i\phi_1}[(1-e^{i\phi_2})\psi_a\psi_a^* - 1]\right|^2 \\ &= 2(1-\cos\phi_2)(|\psi_a|^4 - |\psi_a|^2) + 1 \ , \end{aligned} \tag{26}$$

and

$$\begin{aligned} |b|^2 &= \left|(1-e^{i\phi_2})\psi_a\psi_{a_\perp}^*\right|^2 \\ &= 2(1-\cos\phi_2)(|\psi_a|^2 - |\psi_a|^4) \ , \end{aligned} \tag{27}$$

where b is the quantum transition amplitude (we took its norm in the analyses) and $|b|^2$ is the quantum transition probability[5,6]. Because of the symmetry, the transition probability describes the probability of the transition from $|a\rangle$ to $|a_\perp\rangle$ or the transition from $|a_\perp\rangle$ to $|a\rangle$, measuring the effect of learning from outcome in our system.

Quantum distances

In classical probability theory, if we consider a probability distribution as a state, the distinguishability of states can be computed using the distance measures. Similarly, we can characterize the distinguishability of superposition states using quantum distances. There are several distance measures commonly used in quantum information, such as the trace distance[2], the Hilbert-Schmidt distance[7], the Bures distance[2], and the Hellinger distance[8]. Though different distance measures

have different meanings in physics, such as the geometrical or statistical explanations[9,10], they are equivalent in our system because of the pure state representation of internal states. We use the trace distance for the analyses, and the choice of distance measures will not affect the results in our paper.

Let us consider two superposition states $|\psi_1\rangle$ and $|\psi_2\rangle$, which refer to $|\psi(t)\rangle$ and $|\psi(t+1)\rangle$ in our case. For any superposition state $|\psi\rangle$, its density operator is defined as $\hat{\rho} = |\psi\rangle \langle\psi|$. Correspondingly, we have two density operators $\hat{\rho}_1$ and $\hat{\rho}_2$. Moreover, tr is the trace operator and $\sqrt{\cdot}$ is the principal square root of a positive-semidefinite matrix. The trace distance is then half of the trace norm of the difference of the matrices (the Schatten norm for $p = 1$):

$$\begin{aligned} D_{\mathrm{tr}}(\hat{\rho}_1, \hat{\rho}_2) &= \frac{1}{2}||\hat{\rho}_1 - \hat{\rho}_2||_1 \\ &= \frac{1}{2}\,\mathrm{tr}\,\sqrt{(\hat{\rho}_1 - \hat{\rho}_2)^\dagger(\hat{\rho}_1 - \hat{\rho}_2)}\,. \end{aligned} \tag{28}$$

Relations between quantum transition amplitudes and quantum distances

We define the formula $\hat{U}_{\mathrm{G}} = \hat{Q}_2(\phi_2)\hat{Q}_1(\phi_1)$ as the primal QRL and the equality $\hat{\tilde{U}}_{\mathrm{G}} = \hat{Q}_2(\phi_1)\hat{Q}_1(\phi_2)$ as the dual QRL. The final probability of the chosen action is symmetric about $\phi_1 = \phi_2$, which means we cannot tell the primal QRL from the dual QRL only by projected probabilities, though they have different effects on the state vectors.

It can be numerically verified that the quantum transition amplitude $|b|$ in the primal QRL is the quantum distance D in the dual QRL, while the proof details will be presented in another theoretical paper (Ji-An, L. et al., unpublished manuscript). Similarly, the quantum distance D in the primal QRL is the quantum transition amplitude $|b|$ in the dual QRL. Though the transition amplitude and the quantum distance have different definitions and meanings, they are not distin-

guishable at the projected probability level. We then define the generalized quantum distance as the geometric average of these two signals, combining the information from them.

Uncertainty

Uncertainty here is computed using the Shannon's entropy:

$$\text{uncertainty}(t) = -\sum_{j=1}^{4} \Pr(\delta_j(t) = 1) \log \Pr(\delta_j(t) = 1) \ . \tag{29}$$

Here, uncertainty(t) means the uncertainty for the task of one subject in trial t computed by one model and $\Pr(\delta_j(t) = 1)$ represents the probability of choosing deck j in trial t.

Supplementary Results

I - The Decay rule is better than the Delta rule with the goodness-of-fit but worse with the simulation method

In all groups, when the one-term Decay (EVLDecayTDC, EVLDecayTIC, PVLDecayTDC, PVLDecayTIC) and the one-term Delta models (EVLDeltaTDC, EVLDeltaTIC, PVLDeltaTDC, PVLDeltaTIC) were compared, the one-term Decay model was superior than the one-term Delta model with one-step-ahead predictions. However, this superiority was reversed when the total choice sequences were simulated without using past choices (Figure 3), as reported in former papers[11,12].

II - The QSPP model performs significantly better than other models

We also did Wilcoxon signed rank tests between the QSPP model and other CRL models with false discovery rate (FDR) correction of linear step-up procedure[13]. The QSPP model has smaller AICc values ($p < 0.001$ for all), smaller BIC values ($p = 0.685$ for VPPDecayTDC and $p = 0.007$ for VPPDecayTIC in the control group, and $p < 0.001$ for others), and smaller MSE values ($p < 0.001$ for all).

III - Similarities in parameters between the QSPP and the CRL models

The QSPP model adopted a time independent rule (TIC) for the perseverance part. Hence, we compared the reinforcement learning weight w in the QSPP model with those in the VPPDeltaTIC and the VPPDecayTIC models. A smaller w means that the perseverance plays a larger role in the models. We found that there is a larger correlation between the QSPP w and the VPPDecayTIC w, compared with the correlation between the QSPP w and the VPPDeltaTIC w in the control and smoking groups (Figure S9a-b), showing a larger similarity of the role that the perseverance plays in the QSPP and the VPPDecayTIC models. We also found that, in the control and the smoking groups, the increase of model fit ability (log likelihood) by adding the perseverance into the one-term models (QSL, PVLDecayTIC and PVLDeltaTIC) has a stronger negative correlation with w in the QSPP and the VPPDecayTIC models than w in the VPPDeltaTIC model (Figure S10a-b). This means that if the perseverance cannot help model fitting, it tends to have a small impact on model behavior. The larger similarity of the perseverance term in the QSPP and the VPPDecayTIC models indicates that the VPPDecayTIC model is a better counterpart of the QSPP model in our fMRI analyses. One possible reason why the correlation between w and the increase of model fit

in the VPPDeltaTIC model is near zero in our groups is that the PVLDeltaTIC model showed bad performance for most subjects, while the incorporation of the perseverance term provides a general improvement on model fitting.

IV - The basic reinforcement learning signals were replicated in the VPPDecayTIC model

To demonstrate the rationality of fMRI analysis comparing the QSPP and the VPPDecayTIC models, we also replicated some basic reinforcement learning signals in the VPPDecayTIC model, consistent with former studies[14–16]. We found that the reward prediction error signal was positively related to the activation in the striatum (in the caudate and extended into putamen) and the current action value signal was positively related to the activation in the ventromedial prefrontal cortex (vmPFC) (Figure S8). It was not suitable to test these basic reinforcement learning signals in the QSPP model because the QRL models do not assume action value or prediction error per se.

V - The quantum learning variables showed consistent results

In the control group, the quantum transition amplitude significantly activated the medial frontal gyrus (MeFG)/anterior cingulate cortex (ACC), 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 (Figure S11a). The 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 (Figure S11b). These results were consistent with the generalized quantum distance (Figure 6 in the main text, also see the Supplementary Table S2), showing the

rationality of the definition of the generalized quantum distance.

VI - Uncertainty and uncertainty modulation are represented at the neural level

In the control group, uncertainty from the QSPP model positively related to the activation in a set of regions, including the right ITG/fusiform gyrus, the left fusiform gyrus, the left MeFG, and the left MiFG (see Figure S12 and the Supplementary Table S3). These brain regions might play a role in the representation of uncertainty computed by the QSPP model. The corresponding regions of VPPDecayTIC were similar to that of the QSPP model and there was no significant difference between them.

We further studied the interaction of uncertainty and the generalized quantum distance, and found that it activated the PCC, the vmPFC/ACC, and other areas (Figure S13). These results indicate that uncertainty modulates the main computation of the QRL models.

Supplementary Discussion

I - Uncertainty is represented at the neural level

Uncertainty decreases with time, meaning the subject's uncertainty about the goodness of decks (the first-order uncertainty for each deck) diminishes gradually. Therefore, it can be considered as the second-order uncertainty, or estimation uncertainty[17,18]. Our uncertainty results were consistent with former studies, like the middle frontal gyrus (MiFG)[17,18]. These regions might serve to provide second-order uncertainty for uncertainty-driven exploration[19] and are also involved in

other kinds of uncertainty and ambiguity[18,20].

Though uncertainty is not a direct component in the QRL formulas, it plays a significant role in understanding the QRL algorithms cognitively. First, uncertainty is important to general reinforcement learning and value-based decision making. Uncertainty adjusts the weights of different experiences and optimizes learning rate during value-based decision making according to the agent's environment[20,21]. It also drives exploration and belief updating[19,22] and is involved in arbitration between different reinforcement learning systems[23]. Second, uncertainty is especially important to the IGT. Indeed, the IGT was first introduced to explore decision making under uncertainty[24]: uncertainty fosters learning through somatic markers. In addition, uncertainty modulates learning of prediction errors in the IGT[16]. Third, uncertainty is also important to QRL because it modulates the general reinforcement learning process, including QRL-related learning. Our results of interaction of uncertainty and penalty/reward reflect the influence of learning on internal states. We further studied the interaction of uncertainty and the generalized quantum distance and found that it activated the PCC, the vmPFC/ACC, and other areas. The PCC and the vmPFC/ACC were previously reported to be activated during updating the expectancies of decks and modulated by uncertainty in the IGT[16]. These results indicate that uncertainty modulates the QRL computation and the learning from outcomes in the IGT.

II - The QRL algorithm can function at the neural level with the assumption of classical recurrent neural networks

We have discussed that the medial frontal gyrus (MeFG) and other regions related to the internal-state-based variables, might work as a network in a quantum-like manner, integrating the outcomes

to update internal states in the QRL models. We refer to such a network as a quantum-processing network.

According to the theoretical existence proof[25], quantum algorithms can function if there exist three kinds of classical recurrent neural networks: the unitary evolution network, the choice probability network, and the state reduction network. The former two are implied in our QRL models. We can assume a possible mechanism of how QRL could be implemented in the quantum-processing network.

First, the superposition state $|\psi(t)\rangle$ is kept persistently in the unitary evolution network and sent to the choice probability network in each trial. The quantum-processing network might altogether represent the internal states.

Second, the choice probability network calculates the probability from the probability amplitude and generates an action selection. Such processes might rely on the MeFG and the connections between the MeFG and motor control areas (such as the precentral gyrus) involving motor planning and preparation.

Third, the unitary evolution network receives external outcomes (might be represented in the OFC), performs Grover iteration operator $\hat{U}_{\mathrm{G}}$ on $|\psi(t)\rangle$, and generates $|\psi(t+1)\rangle$ for next trial.

The function of the unitary evolution network is based on neural synchronization and oscillation which are important in cortical computation[26] and high-level cognitive functions like learning and memory[27].

III - Notes on the groups studied

We did not match the control group with the smoking group in the present study in the aspects of sex, age, and education years. One main reason was that the two groups were collected separately and we did not particularly aim at studying the mechanism of smoking addiction. Therefore, controlling these factors was not necessary. In addition, it was rather difficult to recruit female smokers in China, since 52.9% of Chinese adult men smoke, compared with the 2.4% of adult women who smoke[28].

Supplementary Figures

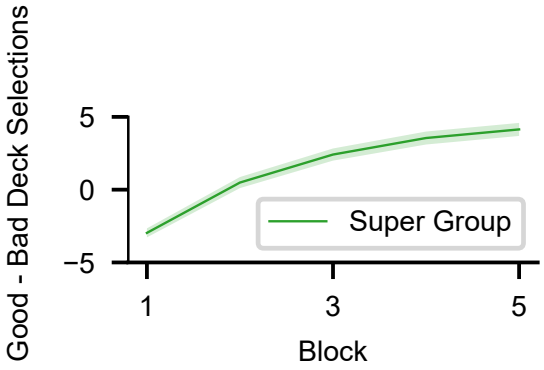

Figure S1: The task performance in the super group, showing proportion of good minus bad deck selections in 20-trial blocks. The shaded regions indicate the standard error.

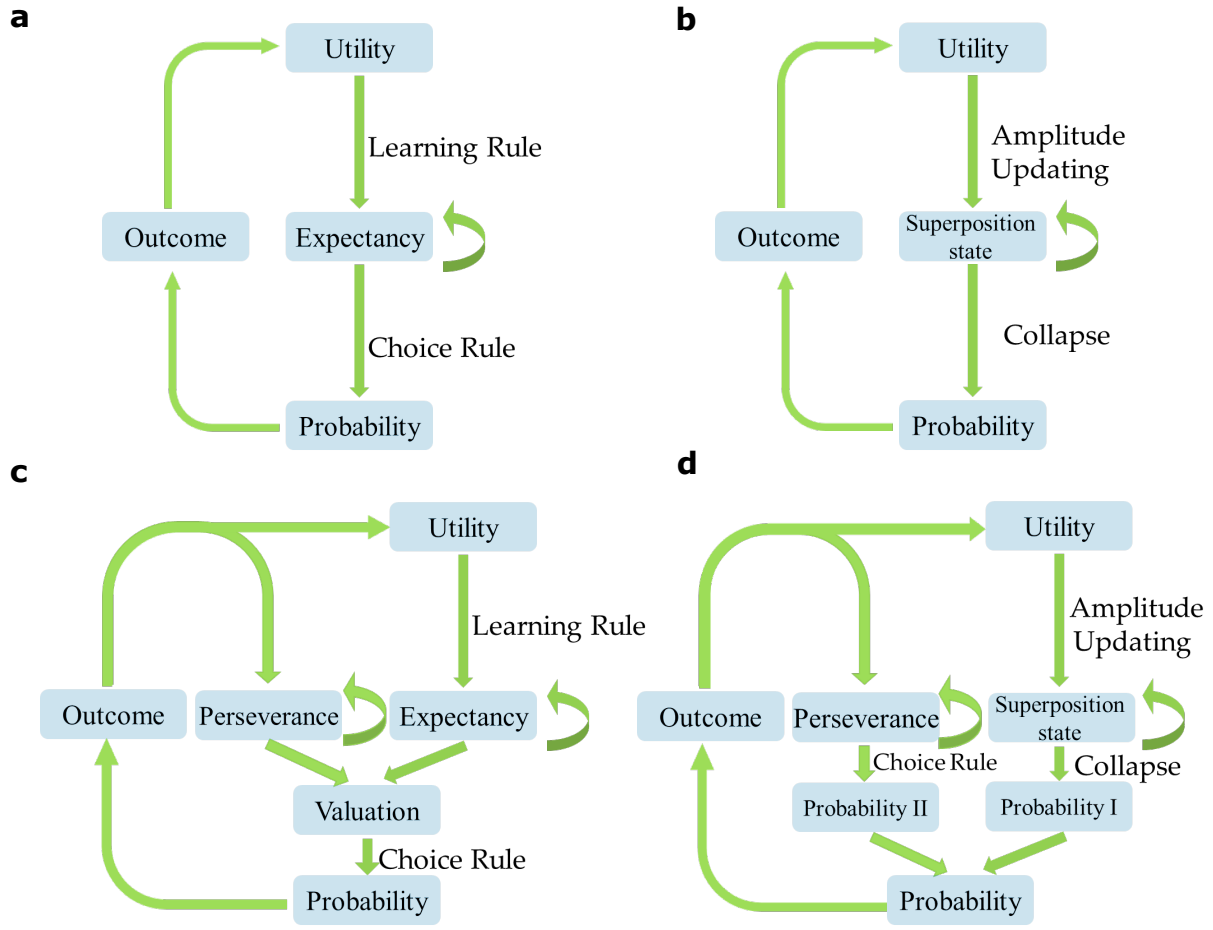

Figure S2: The diagrams for the architecture of reinforcement learning models. **a**, one-term classical reinforcement learning (CRL) models, containing three stages: (1) evaluate the outcome; (2) update the expectancy; and, (3) make a choice. **b**, one-term quantum reinforcement learning (QRL) model (Quantum-Superposition-state-Learning, QSL). **c**, two-term classical reinforcement learning (CRL) models. **d**, two-term quantum reinforcement learning (QRL) model (Quantum-Superposition-state-Plus-Perseverance, QSPP). Model details and comparisons should be referred to the main text and Supplementary Methods: "Computational modeling".

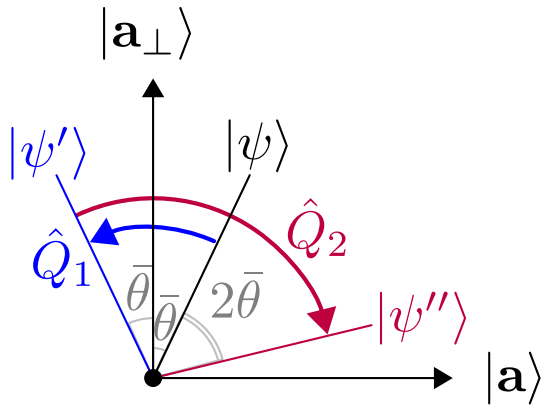

Figure S3: Geometric explanation of simple $\hat{U}_{\mathrm{G}}$ for fixed learning factors. $\hat{Q}_1$ rotates $|\psi\rangle$ into $|\psi'\rangle$ (blue), and $\hat{Q}_2$ rotates $|\psi'\rangle$ into $|\psi''\rangle$ (purple).

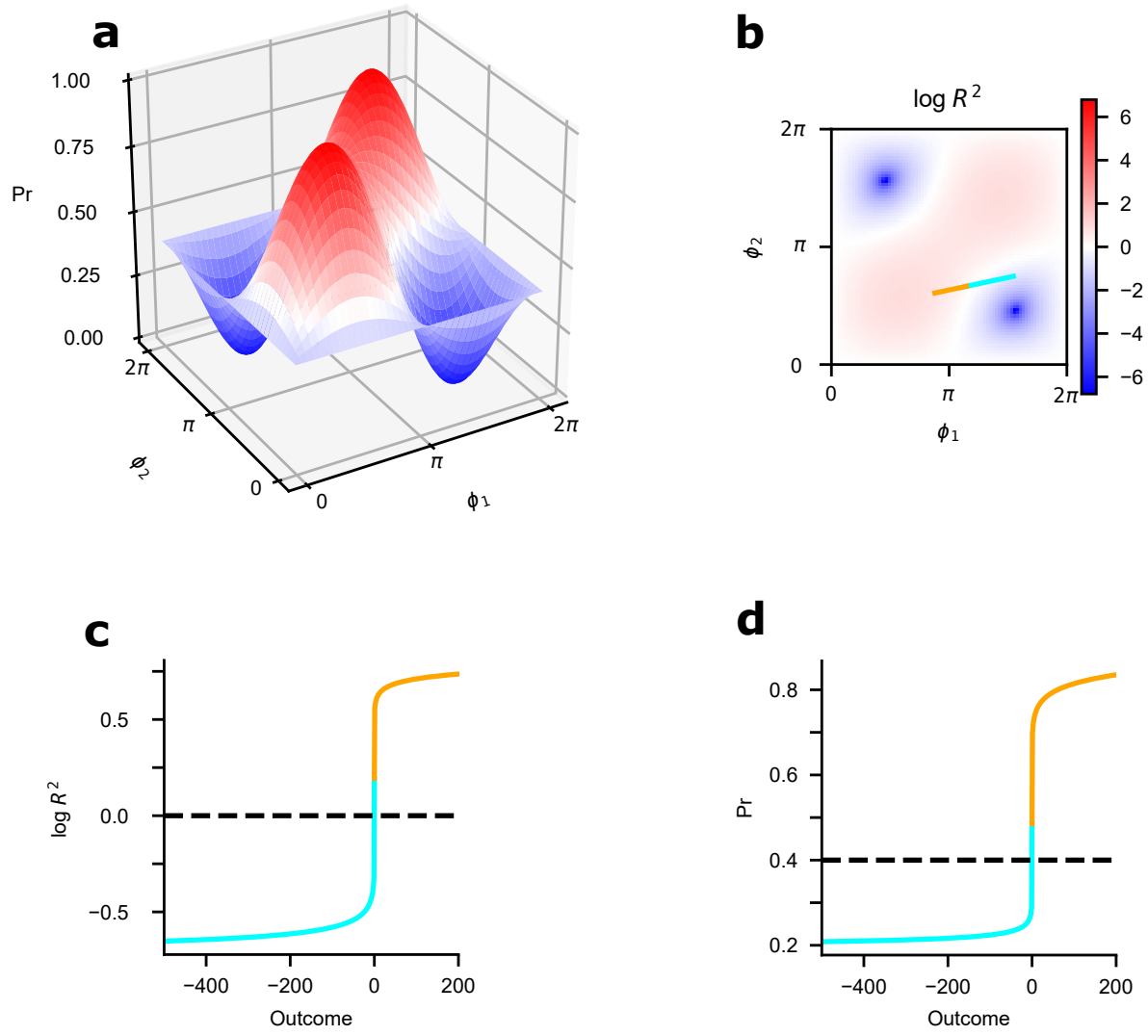

Figure S4: The algebraic visualization of the Grover iteration, showing the case when the probability of the current action is $p = 0.4$. **a**, the new probability p_{new} of the current action as a function of ϕ_1 and ϕ_2. **b**, $LR = \log p_{\text{new}}/p$ as a function of ϕ_1 and ϕ_2. A LR value larger than zero means reinforcement (red) and smaller than zero means penalization (blue). The mapping from $u(t)$ to (ϕ_1, ϕ_2) (determined by the parameters (η, b_1, b_2) from one subject) then delineates a line with slope $\tan \pi\eta$ moving from the starting point (b_1, b_2) (cyan line: the outcome is smaller than zero; orange line: the outcome is larger than zero). **c**, LR as a function of the outcome. **d**, p_{new} as a function of the outcome.

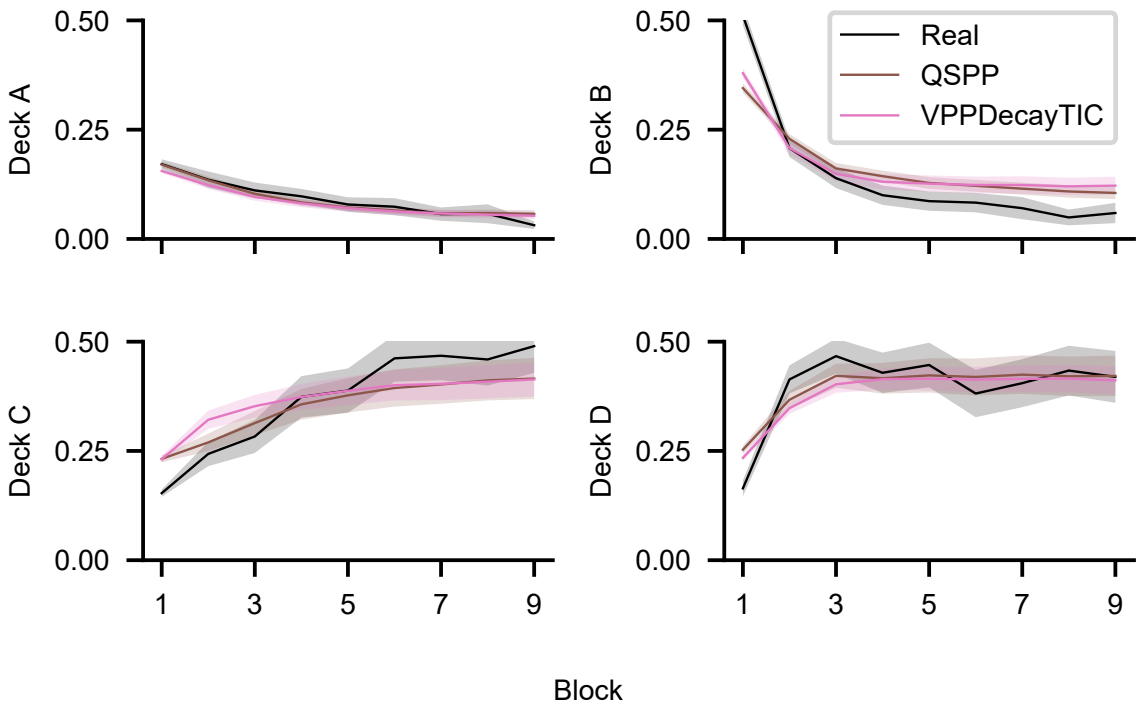

Figure S5: The proportion of each deck selection (EDS) computed by the QSPP and the VPPDecayTIC models in the control group (averaging over subjects). The shaded regions indicate the standard error.

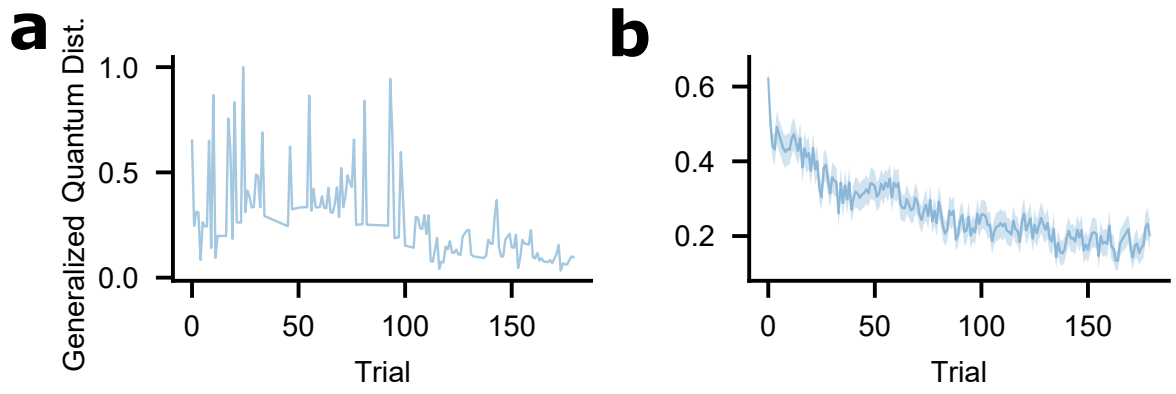

Figure S6: The generalized quantum distance provided by the QSPP model. **a**, the generalized quantum distance of one example subject in the control group. **b**, the generalized quantum distance averaging over subjects in the control group. All shaded regions indicate the standard error.

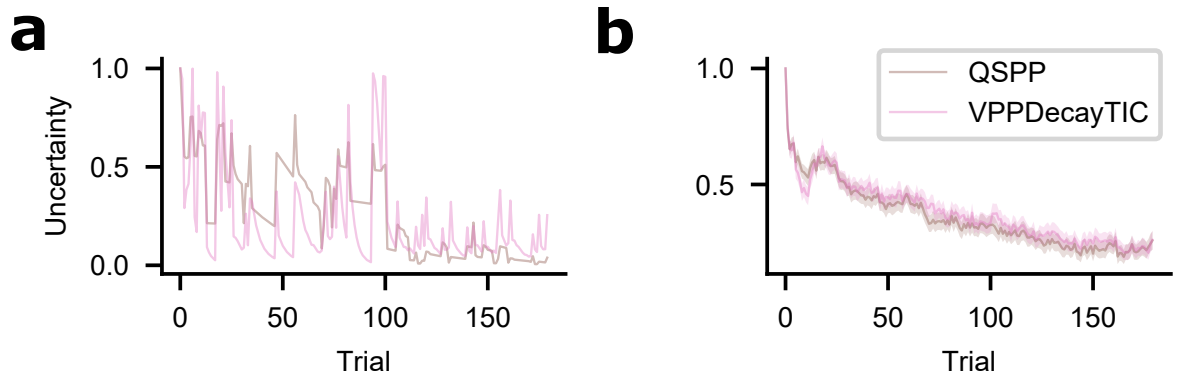

Figure S7: Uncertainty provided by the QSPP and the VPPDecayTIC models. **a**, uncertainty of one example subject in the control group. **b**, uncertainty averaging over subjects in the control group. All shaded regions indicate the standard error.

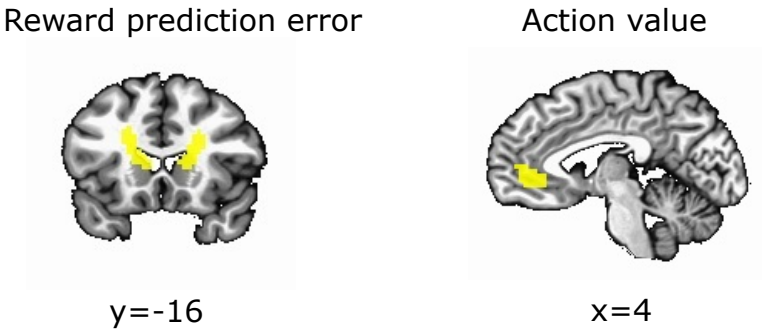

Figure S8: fMRI results of the reward prediction error and the current action value signal in the VPPDecayTIC model in the control group. Reward prediction error signal was related to the activation in the striatum (in the caudate and extended into putamen) and the current action value signal was related to the activation in the ventromedial prefrontal cortex (vmPFC). All coordinates are plotted in RAI (DICOM) order (-right +left, -anterior +posterior, -inferior +superior).

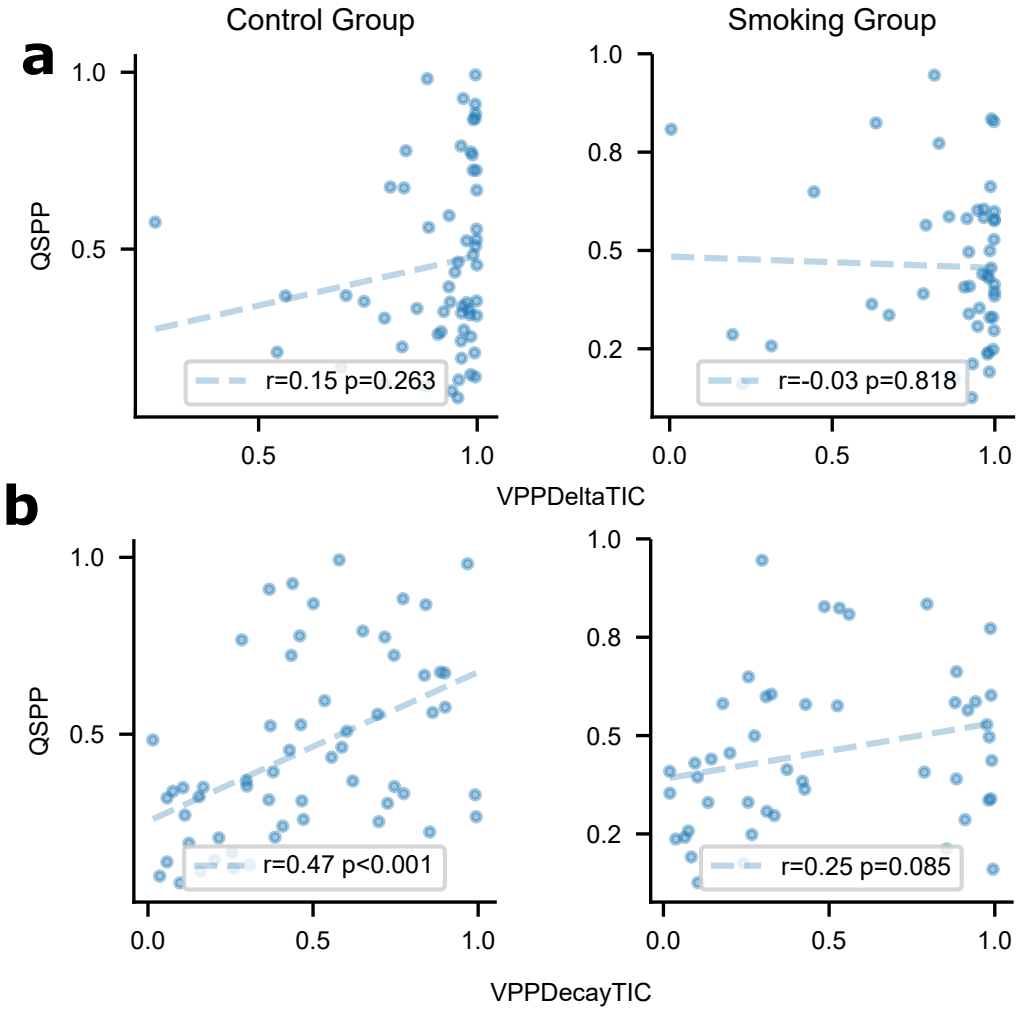

Figure S9: Correlation between the reinforcement learning weight w in the QSPP and in the VPPDeltaTIC models as well as between w in the QSPP and in the VPPDecayTIC models in the control and the smoking groups. **a**, the correlation between the QSPP and the VPPDeltaTIC models (first row: scatter plot of w; second row: correlation of four perseverance parameters). **b**, the correlation between the QSPP and the VPPDecayTIC models.

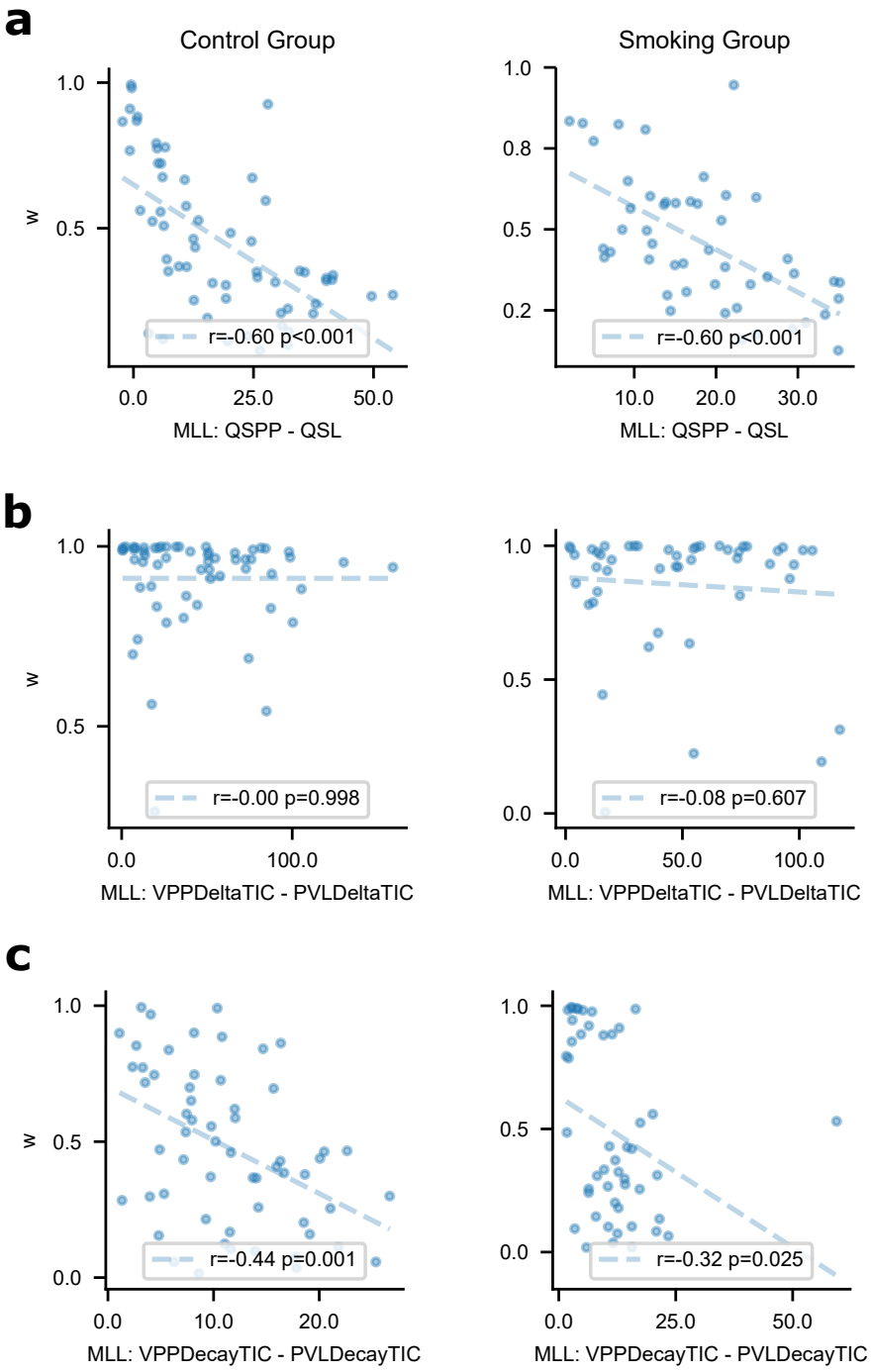

Figure S10: Correlation between the reinforcement learning weight w and the increase of model fit (log likelihood) by incorporating the perseverance term in the QSL, the PVLDeltaTIC and the PVLDecayTIC models in the control and the smoking groups. **a**, the case for the QSPP and the QSL models. **b**, the case for the VPPDeltaTIC and the PVLDeltaTIC models. **c**, the case for the VPPDecayTIC and the PVLDecayTIC models.

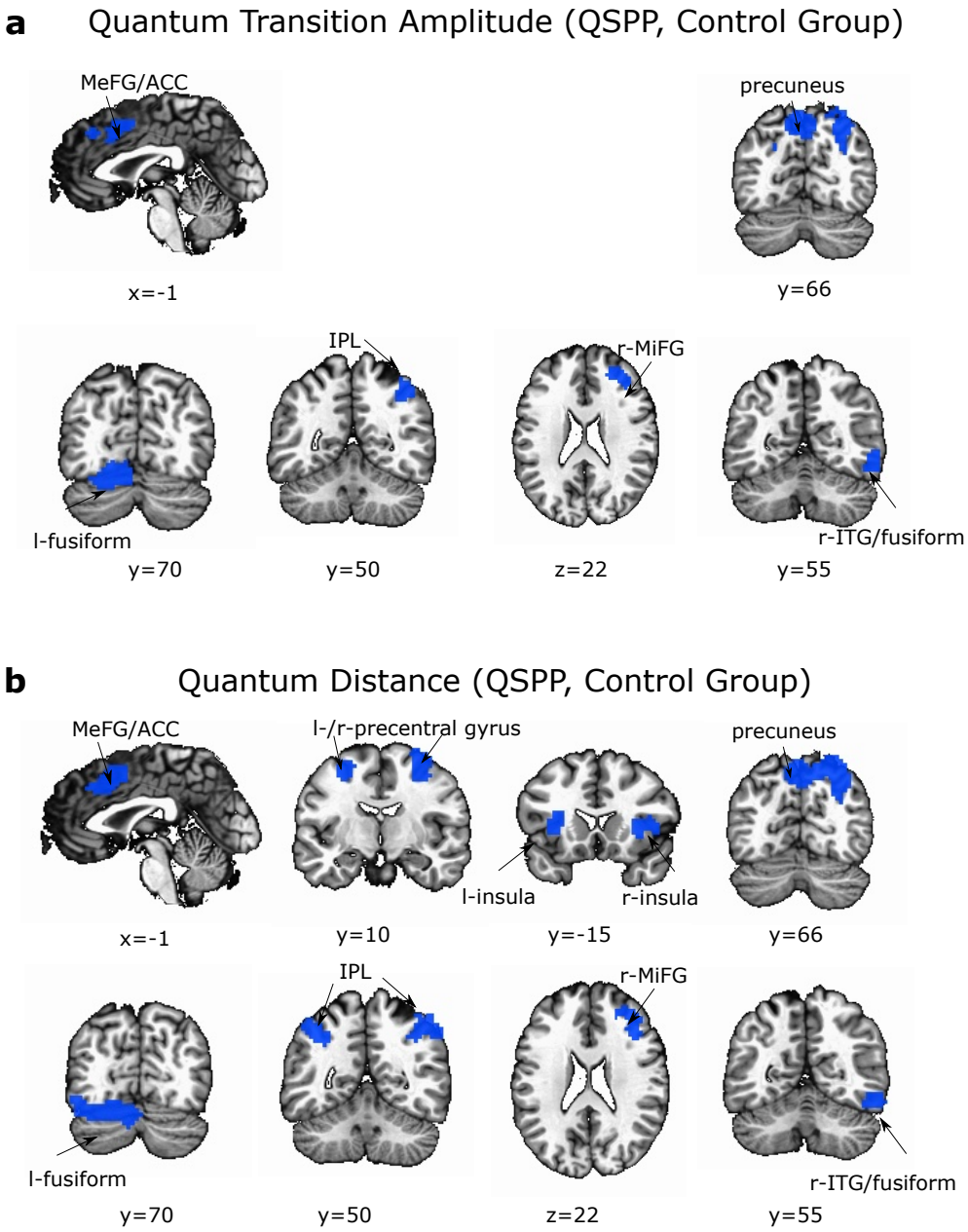

Figure S11: Activity related to the quantum transition amplitude and the quantum distance in the control group. The quantum transition amplitude significantly activated the medial frontal gyrus (MeFG)/anterior cingulate cortex (ACC), the precuneus, the left fusiform gyrus, the inferior parietal lobule (IPL), the middle frontal gyrus (MiFG), and the right inferior temporal gyrus (ITG)/fusiform gyrus. The 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 are plotted in RAI (DICOM) order (-right +left, -anterior +posterior, -inferior +superior).

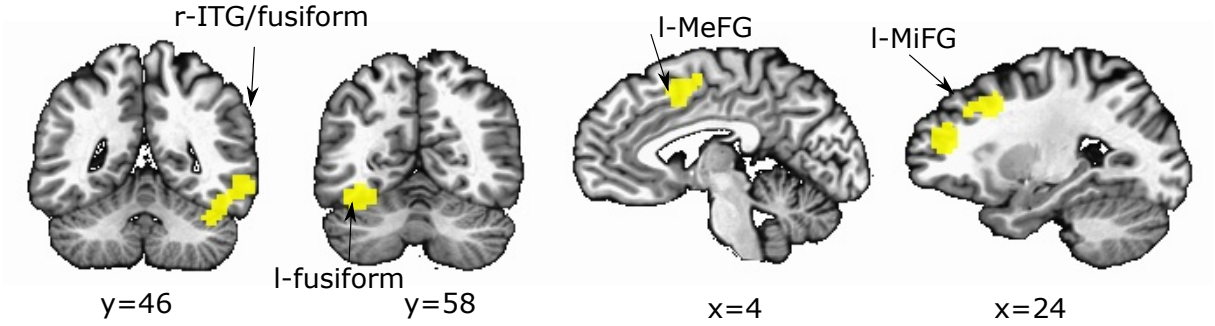

Figure S12: The uncertainty-related activity in the control group. Uncertainty computed by the QSPP model positively related to these regions, including the right inferior temporal gyrus (ITG)/fusiform gyrus, the left fusiform gyrus, the left medial frontal gyrus (MeFG), and the left middle frontal gyrus (MiFG). All coordinates are plotted in RAI (DICOM) order (-right +left, -anterior +posterior, -inferior +superior).

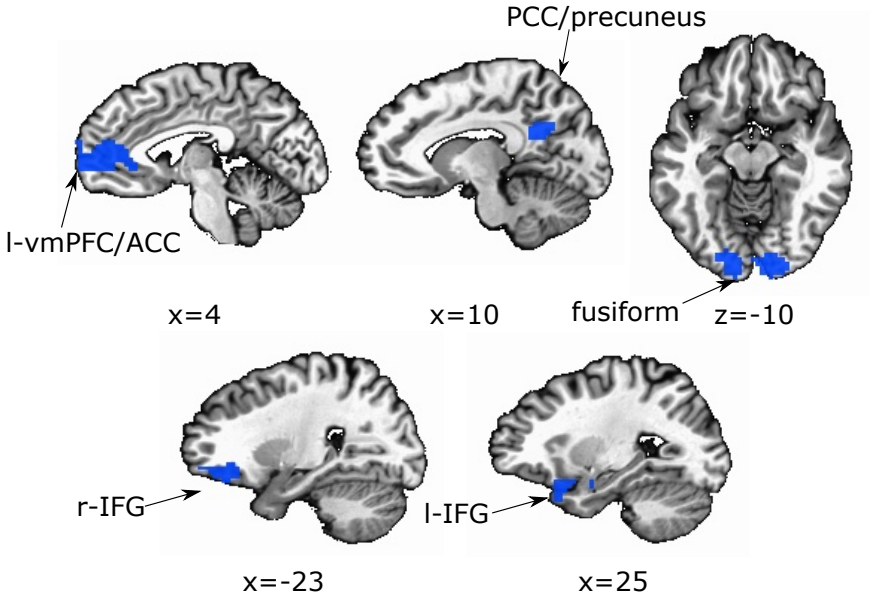

Figure S13: The activity related to the uncertainty $\times$ generalized quantum distance interaction in the control group. The interaction significantly activated the anterior cingulate cortex (ACC)/ventral medial prefrontal cortex (vmPFC), the posterior cingulate cortex (PCC)/precuneus, the fusiform gyrus, and the inferior frontal gyrus (IFG). All coordinates are plotted in RAI (DICOM) order (-right +left, -anterior +posterior, -inferior +superior).

Supplementary Tables

Table S1: From classical to quantum models.

Framework	CRL	QRL	Original QRL
Action representation	Set-theoretic a_1, a_2, a_3, a_4	Geometric $\|a_1\rangle$, $\|a_2\rangle$, $\|a_3\rangle$, $\|a_4\rangle$	Geometric
State representation	Probability distribution $p_j(t)$	Wave function $\|\psi\rangle = \sum_{j=1}^{4} \psi_j \|a_j\rangle$	Wave function
Action valuation	Value function	Wave function	Value function
Action selection	Boltzmann exploration	Superposition state collapse	Superposition state collapse
Outcome evaluation	Utility	Utility	Outcome
Learning rule	Value or perseverance updating	Grover iteration	Value updating and Grover iteration

Table S2: Brain areas correlated with the generalized quantum distance computed by the QSPP model in the control group.

Region	x (mm)*	y (mm)	z (mm)	Extent (voxels)
MeFG/ACC	-2	-5	44	731
L precentral	28	10	53	88
R Precentral	-31	13	59	466
L Insula	28	-16	11	84
R Insula	-31	-13	11	162
Precuneus	4	70	50	1185
L fusiform	19	79	-18	1865
R ITG/fusiform	-52	58	-12	267
L IPL	31	52	38	85
R IPL	-49	40	47	593
L MiFG	40	-34	23	69
R MiFG	-31	-37	29	638

* Coordinates are in Talairach space and RAI (DICOM) order and correspond to the peak of the cluster.

Table S3: Brain areas correlated with the uncertainty computed by the QSPP model in the control group.

Region	x (mm)*	y (mm)	z (mm)	Extent (voxels)
R ITG/fusiform	-55	46	-9	308
L Fusiform	34	58	-18	251
L MiFG	28	-16	38	253
L MeFG	4	-7	44	68

* Coordinates are in Talairach space and RAI (DICOM) order and correspond to the peak of the cluster.

Table S4: Brain areas correlated with the uncertainty × penalty interaction.

Group	Region	x (mm)*	y (mm)	z (mm)	Extent (voxels)
Control group	L MeFG/ACC	19	-52	23	852
	L MiTG/STG	55	10	-12	748
	R MiTG	-61	13	-6	273
	R MeFG	-13	-28	38	99
	L OFC	28	-40	0	134
	L MiTG/angular/PCC/precuneus	43	61	29	1393
Smoking group	L MeFG	10	-28	47	72
	L MiTG/STG	58	10	8	226
	L IPL	31	40	56	144

* Coordinates are in Talairach space and RAI (DICOM) order and correspond to the peak of the cluster.

Table S5: Brain areas correlated with the uncertainty $\times$ reward interaction in the control group.

Region	x (mm)*	y (mm)	z (mm)	Extent (voxels)
L MiTG/STG	61	40	2	86
L MiTG/angular	40	70	32	244
L MeFG	16	-31	41	53**

* Coordinates are in Talairach space and RAI (DICOM) order and correspond to the peak of the cluster.

** Nearly significant.

Table S6: Brain areas correlated with the uncertainty × generalized quantum distance interaction computed by the QSPP model in the control group.

Region	x (mm)*	y (mm)	z (mm)	Extent (voxels)
L vmPFC/ACC	4	-52	2	363
R Fusiform	-16	88	-9	171
L Fusiform	22	88	-15	169
L Culmen	13	49	-12	166
L IFG	25	-16	-18	129
L PCC/precuneus	10	55	26	91
R IFG	-22	-34	-6	77

* Coordinates are in Talairach space and RAI (DICOM) order and correspond to the peak of the cluster.

References

[1] Daoyi Dong, Chunlin Chen, Hanxiong Li, and Tzyh-Jong Tarn. Quantum reinforcement learning. *IEEE Trans. Syst., Man, Cybern., Part B (Cybern.)*, 38(5):1207–1220, 2008.

[2] Michael A Nielsen and Isaac L Chuang. *Quantum Computation and Quantum Information*. Cambridge University Press, 2010.

[3] Simon L Altmann. *Rotations, quaternions, and double groups*. Courier Corporation, 2005.

[4] Pegah Fakhari, Karthikeyan Rajagopal, S. N. Balakrishnan, and J. R. Busemeyer. Quantum inspired reinforcement learning in changing environment. *New Math. Nat. Comput.*, 9(03): 273–294, 2013.

[5] Asher Peres. *Quantum theory: concepts and methods*, volume 57. Springer Science & Business Media, 2006.

[6] Gregor Tanner. Unitary-stochastic matrix ensembles and spectral statistics. *J. Phys. A*, 34 (41):8485, 2001.

[7] Masanao Ozawa. Entanglement measures and the Hilbert–Schmidt distance. *Phys. Lett. A*, 268(3):158–160, 2000.

[8] Shunlong Luo and Qiang Zhang. Informational distance on quantum-state space. *Phys. Rev. A*, 69(3):032106, 2004.

[9] Jerzy Dajka, Jerzy Łuczka, and Peter Hänggi. Distance between quantum states in the presence of initial qubit-environment correlations: A comparative study. *Phys. Rev. A*, 84(3): 032120, 2011.

[10] Ingemar Bengtsson and Karol Życzkowski. *Geometry of quantum states: an introduction to quantum entanglement*. Cambridge university press, 2017.

[11] Eldad Yechiam and Jerome R Busemeyer. Comparison of basic assumptions embedded in learning models for experience-based decision making. *Psychon. Bull. Rev.*, 12(3):387–402, 2005.

[12] Woo-Young Ahn, Jerome R Busemeyer, Eric-Jan Wagenmakers, and Julie C Stout. Comparison of decision learning models using the generalization criterion method. *Cogn. Sci.*, 32(8): 1376–1402, 2008.

[13] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.

[14] Yuan Chang Leong, Angela Radulescu, Reka Daniel, Vivian DeWoskin, and Yael Niv. Dynamic interaction between reinforcement learning and attention in multidimensional environments. *Neuron*, 93(2):451–463, 2017.

[15] John P O'doherty, Jeffrey Cockburn, and Wolfgang M Pauli. Learning, reward, and decision making. *Annu. Rev. Psychol.*, 68:73–100, 2017.

[16] Ying Wang, Ning Ma, Xiaosong He, Nan Li, Zhengde Wei, Lizhuang Yang, Rujing Zha, Long Han, Xiaoming Li, Daren Zhang, Ying Liu, and Xiaochu Zhang. Neural substrates of updating the prediction through prediction error during decision making. *NeuroImage*, 157: 1–12, 2017.

[17] Elise Payzan-LeNestour, Simon Dunne, Peter Bossaerts, and John P ODoherty. The neural representation of unexpected uncertainty during value-based decision making. *Neuron*, 79 (1):191–201, 2013.

[18] Dominik R Bach, Oliver Hulme, William D Penny, and Raymond J Dolan. The known unknowns: neural representation of second-order uncertainty, and ambiguity. *J. Neurosci.*, 31(13):4811–4820, 2011.

[19] David Badre, Bradley B Doll, Nicole M Long, and Michael J Frank. Rostrolateral prefrontal cortex and individual differences in uncertainty-driven exploration. *Neuron*, 73(3):595–607, 2012.

[20] Timothy E J Behrens, Mark W Woolrich, Mark E Walton, and Matthew F S Rushworth. Learning the value of information in an uncertain world. *Nat. Neurosci.*, 10(9):1214–1221, 2007.

[21] Ming Hsu, Meghana Bhatt, Ralph Adolphs, Daniel Tranel, and Colin F Camerer. Neural systems responding to degrees of uncertainty in human decision-making. *Science*, 310(5754): 1680–1683, 2005.

[22] Joseph T McGuire, Matthew R Nassar, Joshua I Gold, and Joseph W Kable. Functionally dissociable influences on learning rate in a dynamic environment. *Neuron*, 84(4):870–881, 2014.

[23] Nathaniel D Daw, Yael Niv, and Peter Dayan. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat. Neurosci.*, 8(12):1704–1711, 2005.

[24] Antoine Bechara, Antonio R Damasio, Hanna Damasio, and Steven W Anderson. Insensitivity to future consequences following damage to human prefrontal cortex. *Cognition*, 50(3): 7–15, 1994.

[25] J. R. Busemeyer, P Fakhari, and P Kvam. Neural implementation of operations used in quantum cognition. *Prog. Biophys. Mol. Bio.*, 130:53–60, 2017.

[26] Pascal Fries. Neuronal gamma-band synchronization as a fundamental process in cortical computation. *Annu. Rev. Neurosci.*, 32:209–224, 2009.

[27] Stephen Grossberg and Massimiliano Versace. Spikes, synchrony, and attentive learning by laminar thalamocortical circuits. *Brain Res.*, 1218:278–312, 2008.

[28] Qiang Li, Jason Hsia, and Gonghuan Yang. Prevalence of smoking in China in 2010. *N. Engl. J. Med.*, 364(25):2469–2470, 2011.