HAL
open science

Conflict-free collective stochastic decision making by orbital angular momentum of photons through quantum interference

Takashi Amakasu, Nicolas Chauvet, Guillaume Bachelier, Serge Huant, Ryoichi Horisaki, Makoto Naruse

To cite this version:

Takashi Amakasu, Nicolas Chauvet, Guillaume Bachelier, Serge Huant, Ryoichi Horisaki, et al.. Conflict-free collective stochastic decision making by orbital angular momentum of photons through quantum interference. Scientific Reports, 2021, 11, pp.21117. 10.1038/s41598-021-00493-2 . hal-03376331

HAL Id: hal-03376331

https://hal.science/hal-03376331v1

Submitted on 4 Nov 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Conflict-free collective stochastic decision making by orbital angular momentum of photons through quantum interference

Takashi Amakasu[1,*], Nicolas Chauvet[1,2], Guillaume Bachelier[3], Serge Huant[3], Ryoichi Horisaki[1,2], and Makoto Naruse[1,2,*]

[1]Department of Mathematical Engineering and Information Physics, Faculty of Engineering, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan.
[2]Department of Information Physics and Computing, Graduate School of Information Science and Technology, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan.
[3]Université Grenoble Alpes, CNRS, Institut Néel, 38042, Grenoble, France.
[*]T.A. (email: amakasutakashi@gmail.com), M.N. (email: makoto_naruse@ipc.i.u-tokyo.ac.jp)

arXiv:2107.00877v2 [quant-ph] 18 Sep 2021

ABSTRACT

In recent cross-disciplinary studies involving both optics and computing, single-photon-based decision-making has been demonstrated by utilizing the wave-particle duality of light to solve multi-armed bandit problems. Furthermore, entangled-photon-based decision-making has managed to solve a competitive multi-armed bandit problem in such a way that conflicts of decisions among players are avoided while ensuring equality. However, as these studies are based on the polarization of light, the number of available choices is limited to two, corresponding to two orthogonal polarization states. Here we propose a scalable principle to solve competitive decision-making situations by using the orbital angular momentum of photons based on its high dimensionality, which theoretically allows an unlimited number of arms. Moreover, by extending the Hong-Ou-Mandel effect to more than two states, we theoretically establish an experimental configuration able to generate multi-photon states with orbital angular momentum and conditions that provide conflict-free selections at every turn. We numerically examine total rewards regarding three-armed bandit problems, for which the proposed strategy accomplishes almost the theoretical maximum, which is greater than a conventional mixed strategy intending to realize Nash equilibrium. This is thanks to the quantum interference effect that achieves no-conflict selections, even in the exploring phase to find the best arms.

Introduction

Optics and photonics are expected to play crucial roles in future computing systems[1], making a variety of devices and systems to be intensively studied such as optical fibre-based neuromorphic computing[2], on-chip optical neural networks[3], optical reservoir computing[4], among others. While these works are basically categorized in supervised learning, reinforcement learning is another important branch of artificial intelligence[5]. The Multi-Armed Bandit (MAB) problem is an example of a reinforcement learning situation, which formulates a fundamental issue of decision making in dynamically changing uncertain environments where the target is to find the best selection among many slot machines, also referred to as arms, whose reward probabilities are unknown[6]. In solving MAB problems, exploration actions are necessary to find the best arm, although too much exploration may reduce the final amount of obtained reward from the exploitation. On the opposite, insufficient exploration may lead to miss the best arm. Furthermore, when multiple players are involved, decision conflicts become serious, as they induce congestions and inhibit socially achievable benefits[7,8]. Equality among players is another critical issue, as unfair repartition of outcomes may lead to distrust the system. This whole problem is known as the competitive MAB (CMAB) problem.

In order to solve these complex issues, photonic solutions have been recently considered. For example, the wave-particle duality of single photons has been utilized for the resolution of the two-armed bandit problem[9]. Moreover, Chauvet *et al.* theoretically and experimentally demonstrated that polarization entangled photon pairs provide non-conflict and equality-assured decisions in two-player, two-armed bandit problems[10]. Entangled photon states that allow more than three players while guaranteeing optimal outcome and equal repartition have also been demonstrated[11].

However, since these former principles rely on the polarization of light as the tunable degree of freedom, the number of possible selections or arms is limited to only two, although potential scalability for the single-player MAB is feasible within a tournament-based approach[12]. Therefore, the scalable principle of decision-making has been an important and fundamental issue, especially for multiplayer situations. In this paper, we introduce the use of the orbital angular momentum (OAM) of

photons[13],[14] to resolve the scalability issue of photonic decision making, following the concept summarized in Fig. 1.

Photons that carry OAM[13] realize high-dimensional state spaces, only restricted by the precision and accuracy of the generation technique and the transmission medium[15] (Fig. 1a); hence one of the basic ideas of this study is to associate individual selections to different-valued OAM (Fig. 1b). The applications of OAM have progressed in diverse areas ranging from the manipulation of cooled atoms, communications, nonlinear optics, optical solitons, and so on. The high-dimensionality of OAM is particularly attractive for quantum information processing in increasing the dimension of elementary quantum information carriers to go beyond the qubit[16],[17],[18],[19],[20],[21]. Likewise, in the present study, the multi-dimensionality of OAM plays a crucial role in extending the maximum number of arms as well as utilizing the probabilistic attribute of single photons carrying OAM.

Furthermore, to resolve CMAB problems when the number of arms is greater than two, we extend the notion of Hong-Ou-Mandel effect[22] to more than two (OAM) vector states to induce quantum interference. We show that conflicting decisions among two players can be perfectly avoided by the adequate quantum interference design to generate OAM 2-photon states, relying on a coherent photon pair source. In the literature, OAM has been examined from game-theoretic perspectives such as resolving prisoners dilemma[23] and duel game[24]. In the present study, we benefit from quantum interference for non-conflicting decision-making to maximize total rewards, which is similar to the insight gained by quantum game literature. Additionally, in solving CMAB problems with many arms, exploration action is necessary. We numerically examine total rewards regarding three-armed bandit problems where the proposed quantum-interference-based strategy accomplishes nearly theoretical maximum total reward. We confirm that the proposed strategy clearly outperforms conventional ones, including the mixed strategy intending to realize Nash equilibrium[7].

Moreover, equality among players is important in CMAB problems. We demonstrate that equality is perfectly ensured by appropriate quantum interference constructions when the number of arms is three. At the same time, however, we also show that it is unfortunately impossible to accomplish perfect equality in the proposed scheme and with the current hypotheses when the number of arms is equal to or larger than four. Note also that perfect collision avoidance is ensured for any number of arms.

These properties are made possible thanks to the high dimensionalities of OAM for scalability and the quantum interference effect for non-conflict selections even in the exploring phase to find the best arms.

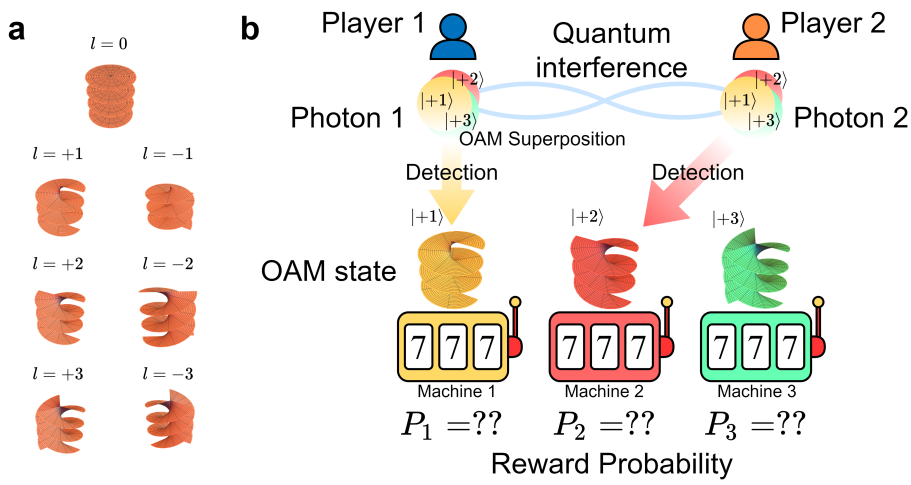


Figure 1. **(a)** Schematic illustration of the phase surface of a beam with OAM. The direction of travel is upward. It is a surface with $|l|$ spirals. The pitch of each spiral is one wavelength divided by $|l|$. **(b)** Stochastic detection of OAM states directly corresponds to a player probabilistically selecting a machine.

Results

Scalable decision maker with OAM

System architecture for solving 1-player K-armed bandit problem.

We first describe the problem under study, which is a stochastic multi-armed bandit problem with rewards following Bernoulli distributions defined as follows. There are K available slot machines (or arms): when the player selects one arm i, the player wins with probability P_i (and receives a fixed reward of 1), or loses with probability $1 - P_i$ (and receives a fixed reward of 0), with i an integer ranging from 1 to K. Let a player choose an arm each time and allow a total of T times, then the goal of the bandit problem is to find out which strategy should be followed to choose arms so that the resultant accumulated outcome is

maximized. When the slot machine with the highest winning probability is known, the best strategy is to draw that specific arm for all T times, but the player initially has no information about the arms. Therefore, exploration actions are required to know the best arm, whereas too much exploration potentially leads to missing a higher total amount of rewards from the best machine.

In the previous work on single-photon decision maker using polarization[9], two orthogonal linear polarizations of photons are associated with two slot machines; that is, horizontal and vertical polarizations correspond to slot machine 1 and 2, respectively. The exploration is physically realized by the probabilistic attribute of photon measurement, whose outcome depends on the direction of the polarization of linearly polarized single photons. Therein, the polarization degree of freedom physically and directly allows specifying the probabilistic selection of slot machines. However, as mentioned in the *Introduction* section, the number of arms is limited to only two, although extendable in a single-player setup to powers of two via a tournament-based approach[12].

The fundamental idea of the present study is to associate the dimension of OAM with the selection of multiple arms, whatever the number of arms. Allen *et al.* have pointed out that a Laguerre-Gaussian (LG) beam has an angular momentum independent from polarization; they have called it OAM to distinguish it from the polarization-dependent spin angular momentum[25]. The spatial mode of a LG beam can be expressed using the near-axis approximation

$$u_l = f_m(\rho, z) e^{il\theta} e^{ikz} \tag{1}$$

where ρ is the distance from the optical axis, θ is the azimuthal angle around the optical axis, z is the coordinate of the propagation direction, f_m is the complex amplitude distribution, and k is the wavenumber. m and l are integer numbers that respectively describe the order of the Laguerre polynomial for the radial distribution and the azimuthal rotation number. In our study, m is fixed at 0, while l takes any integer numbers. Correspondingly, $|l\rangle$ is the state in which there is one photon in the l mode, whose angular momentum is equal to $l\hbar$ where $\hbar$ is Planck's constant divided by 2π. Since the modes with different l are orthogonal to each other, the quantum state can be expressed by a linear superposition, using these modes as a basis. Figure 1a schematically illustrates examples of beams with different l-valued OAM where l is an integer from -3 to 3. Non-zero l beams exhibit spiral isophase spatial distributions. Figure 2 shows a schematic diagram of the proposed system architecture for solving the MAB problem using OAM. Here we illustrate the case where the number of arms is three, but the same principle applies in extending to a larger number of arms.

Conventional laser sources generate beams that do not have orbital angular momentum. Technologically, methods to generate light with OAM from a plane wave or a Gaussian beam include the use of phase plates[26], computer generated holograms (CGH)[27], or mode converters[28,29]. Spatial light modulators (SLMs) are widely utilized for this purpose, as they enable direct and tunable amplitude and/or phase modulation of an incoming light beam[30]. The simplest and the most widely used method is a CGH-based approach implemented with an SLM and a 4f optical setup[15]. In Fig. 2, a photon with a Gaussian spatial profile emitted from a laser is sent to a phase SLM, displaying a CGH pattern to generate OAM states, each carrying a phase factor $e^{il\theta}$ which depends on the azimuthal angle θ and the OAM number l. l could be any integer, but when all generated l are expected to be positive, the output photon is described by the state:

$$SLM(\phi_1, \phi_2, \ldots, \phi_K)\,|0\rangle = \frac{1}{\sqrt{K}} \sum_{l=1}^{K} e^{i\phi_l}\,|+l\rangle \tag{2}$$

where $\phi_1, \phi_2, \ldots, \phi_K$ depict phase changes associated with each OAM with l values being +1, +2, …, and $+K$, respectively, and $|l\rangle$ denotes the photon state with OAM value of l. That is to say, a single photon is emitted from the source system that contain K OAM states with equal probability amplitude.

Meanwhile, a mirror causes flipping of the twisted structure of any given OAM; that is, the function of a beam splitter (BS) in the light propagation is represented by

$$|\Phi\rangle \xrightarrow{\text{1:1 beam splitter}} \frac{1}{\sqrt{2}}\,|\Phi\rangle_{transmitted} + \frac{i}{\sqrt{2}} R\,|\Phi\rangle_{reflected}, \tag{3}$$

where R represents flipping of OAM state, for example, $R|+1\rangle = |-1\rangle$. In the case $K = 3$, we generate a photon state that carries equally $l = +1, +2, +3$ by setting $\phi_1 = \phi_2 = \phi_3 = 0$. That is, the output after SLM is given by $(1/\sqrt{3}) \times (|+1\rangle + |+2\rangle + |+3\rangle)$.

This photon is then transferred to an array of BSs and single photon detection system to examine which l-valued OAM is detected. Among a variety of methods in measuring the OAM of light[31], the system architecture shown in Fig. 2 illustrates a method utilizing a hologram (HG) followed by a zeroth-order extraction system[32]. In practical implementation, a zeroth-order extraction system could be free-space optics with spatial filtering or single-mode optical fibre.

This hologram adds a phase factor of $e^{il_{HG}\theta}$ to the state $|l\rangle$ with OAM l, which results in a transformation $|l\rangle \to |l + l_{HG}\rangle$. After injection into a zeroth-order extraction system, only an $l = 0$ photon propagates in it. In other words, the zeroth-order extraction system acts as a filter to extract the $l = 0$ component only. If the hologram induces a shift of OAM by l_{HG} and a

photon is detected by the subsequent photodetector, the OAM of the incoming photon is identified to be $l = -l_{HG}$. Based on this principle, in the system shown in Fig. 2, three holograms HG1, HG2, HG3 are arranged, which transform $|l\rangle$ into $|l-1\rangle$, $|l-2\rangle$, and $|l-3\rangle$, respectively.

One remark here is that, although multiple BSs and holograms are employed in Fig. 2, more compact realization is indeed possible by, for example, a geometric optical transformation technique[33], which has been extended to more than 50 OAM states[34]. The reason behind the introduction of the measurement architecture shown in Fig. 2 regards the following procedure related to photon detections.

The output light is subjected to attenuators (ATT1, ATT2, ATT3) to control detection probabilities and a zeroth-order extraction system, followed by photodetectors (PD1, PD2, PD3). Based on the filtering by the zeroth-order extraction system, photon detection by PD1, PD2, and PD3 means observing OAM values of 1, 2, and 3, respectively. Photon detection by PD1 immediately means playing slot machine 1. Similarly, PD2 and PD3 are associated with the decision of playing slot machines 2 and 3, respectively. It should be emphasized that in this configuration, a machine is only selected if a photon is detected.

Initially, since the probabilities of the detected photons to be measured by PD1, PD2, and PD3 are all equal to $1/3$, all machines are explored equally. Depending on the obtained results, the attenuation levels by ATT1, ATT2, ATT3 are updated.

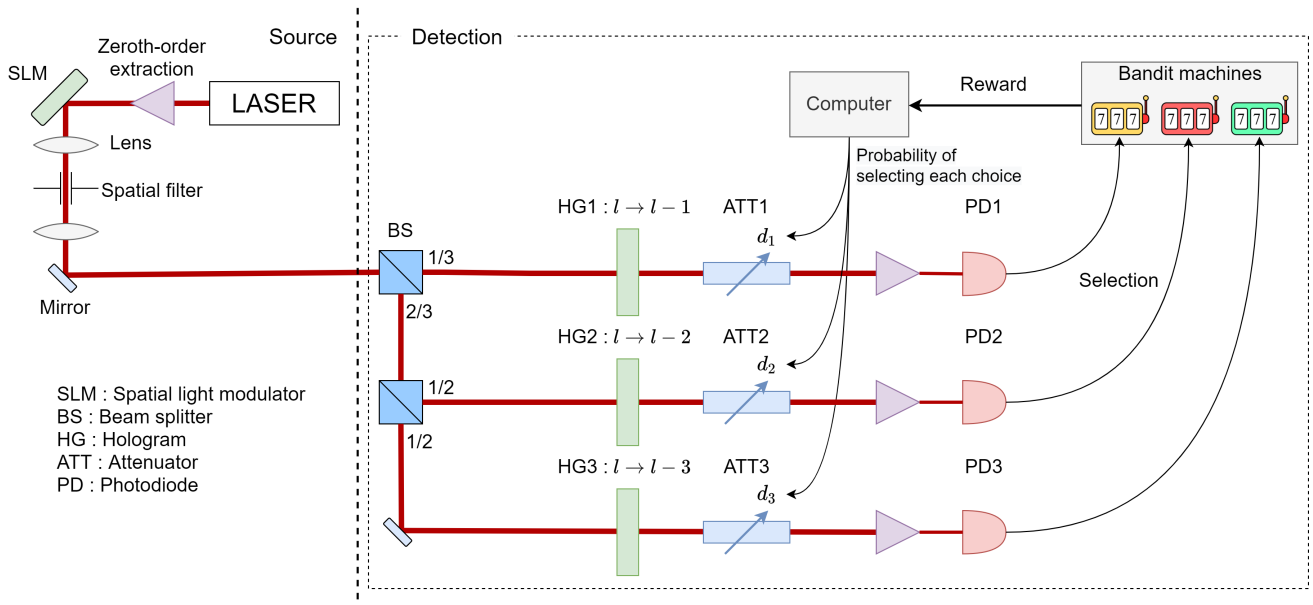


Figure 2. Design of the system architecture for solving single-player 3-armed bandit problems. d_1, d_2, d_3 denote the transmittance of each attenuator.

After a single photon is detected by any photodetector, the selection yields eventual rewards from slot machines, and the results are registered into history $H(t)$. While referring to the history $H(t)$, the next decision is determined by following a certain policy of the player. The softmax policy is one of the most well-known feedback algorithms for the decision, which is also considered to accurately emulate the model of human decision making[35,36]. In the softmax policy, the player selects each machine based on a maximum likelihood estimation of the reward probability $\hat{P}_1(t), \hat{P}_2(t), \ldots, \hat{P}_K(t)$ and the probability of selecting machine i is given by the following equation:

$$s_i(t+1) = \frac{e^{\beta \hat{P}_i(t)}}{\sum_{k=1}^{K} e^{\beta \hat{P}_k(t)}} \tag{4}$$

where β, which is also known as inverse temperature from analogy to statistical mechanics, is a parameter that influences the balance between exploration and exploitation. While optimal parameter β depends on reward probabilities and some methods for tuning β have been proposed[37], this paper, for simplicity, set it to a constant value $\beta = 20$ based on a moderate tuning. The amplitude transmittance of attenuators (ATT1, ATT2, ATT3) are denoted by d_1, d_2, d_3, which are initially all one. These values

are updated after every trial based on:

$$d_i(t) = \sqrt{\frac{s_i(t)}{\max_k s_k(t)}}. \tag{5}$$

In this way, $d_i(t)$ is revised as the time elapses so that the photon detection event is highly likely induced at the photodetector that corresponds to the best slot machine or the highest reward probability machine. For example, if slot machine 1 is the highest reward probability one, the transmittance of ATT1 should be higher while those of ATT2 and ATT3 should become smaller.

Here is a remark about the denominator of the right side of Eq. (5). The probability of detecting state i is proportional to $d_i(t)^2$. Dividing each $d_i(t)^2$ by the same value $\max_k s_k(t)$ does not give any unintended bias to the detection probabilities, but transmission efficiency by the attenuators is kept high. That is, the loss of photons by the attenuators is minimized.

Finally, we discuss one more important remark regarding the architecture for solving the single-player, multi-armed bandit problem shown in Fig. 2. The principle maximizes the detection probability of the OAM state corresponding to the best machine. Actually, instead of reconfiguring the attenuators, we can accomplish the same functionality by reconfiguring the phase pattern displayed at the SLM located on the light source side. Indeed, this alternative way is directly and dynamically utilizing the high-dimensional property of OAM[38]. This architecture, however, imposes a complex arbitration mechanism when we extend the principle to two-player situations in the following. That is, controlling the light source by a single player is indeed feasible, but the source management by two players is non-trivial. Instead, player-specific attenuator control does not impose any global server. For these reasons, we discuss the fundamental architecture shown in Fig. 2.

Simulation results for 1-player 3-armed bandit problem.

Figure 3 summarizes simulation results for the 1-player 3-armed bandit problem with the OAM system following the softmax policy. The solid, dashed, and dashed-dotted curves in Fig. 3a show the time evolution of the selection probability of machine 1, 2, and 3, respectively, when the reward probability of slot machines are given by $[P_1, P_2, P_3] = [0.9, 0.7, 0.1]$. Here the number of repetitions is 1000. We can clearly observe that the probability of selecting the maximum reward probability machine, here machine 1, monotonically increases.

Figure 3b examines the correct decision rate, which is referred to as CDR, defined by the number of selections of the highest reward probability machine over 1000 trials when the reward environment is configured differently. The blue, red, and yellow curves show the time evolution of CDR when the reward environment $[P_1, P_2, P_3]$ is given by $[0.9, 0.7, 0.1]$, $[0.9, 0.5, 0.1]$, and $[0.9, 0.3, 0.1]$, respectively. Here the maximum and minimum reward probabilities are commonly configured. As the difference between the maximum and the second maximum reward probability becomes smaller, the increase of CDR toward unity becomes slow. Nevertheless, we can observe that the monotonic increase of selecting the best machine in Figs. 3a and 3b. Since there is no theoretical limitation regarding the number of OAM states, the system configuration herein can be used for the probabilistic selection among a large number of selections. Note that the softmax policy itself is also scalable.

Solving 2-player 3-armed bandit problem with OAM and quantum interference

System architecture for solving 2-player 3-armed competitive bandit problem with OAM and quantum interference.

This section discusses stochastic selections of arms in the CMAB problem using photon pair OAM quantum states. The system presented in Fig. 2 has been extended to the case of two players (Player A and B) by the architecture represented in Fig. 4. This time, the assumption is that the selection only happens when exactly one photon is detected simultaneously by each player on their photodetectors.

In the source part, a photon pair is created by a nonlinear crystal such as a periodically poled KTP (PPKTP) and then subjected to an interferometer. One of the photon pair is supplied to the Detection A system, and the other goes to the Detection B system. The internal structure of Detection systems is the same as the one-player system depicted in Fig. 2. Thanks to the quantum interference, even though there is no explicit communication between the players, the detection results of the two photons are correlated with each other, as discussed in detail later.

In quantum research using light, it has been common to use quantum states based on properties such as polarization, spatial mode, and phase, but since the discovery of orbital angular momentum, many studies on quantum states using orbital angular momentum of light have been reported[39]. The availability of orbital angular momentum with an infinite number of states is very important in quantum research. In 2001, Mair *et al.* used parametric down conversion (PDC) to study the generation of photon pairs in states with entangled orbital angular momentum[39]. Subsequently, a theoretical study of the change in orbital angular momentum during the PDC process was performed[40], and photon pairs with three entangled orbital angular momentum states were also studied[41]. In the present study, we utilize quantum interference given by an extension of the Hong-Ou-Mandel effect[22].

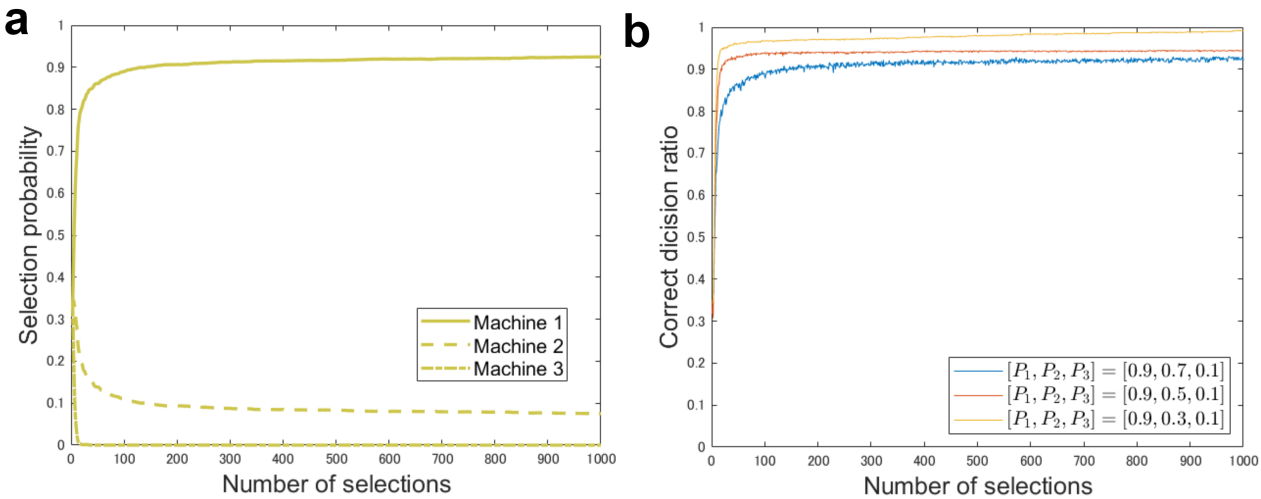


Figure 3. Simulation results for solving 1-player 3-armed bandit problem with OAM system and softmax policy. **(a)** Probability of selecting each machine over 1000 times when $[P_1, P_2, P_3] = [0.9, 0.7, 0.1]$ averaged over 1000 repetitions. **(b)** Correct decision rate over 1000 repetitions in different reward environment.

Generation of OAM photon pair with quantum interference.

Hong-Ou-Mandel effect has been well studied for two identical photons always detected together in the same output path when they enter into a 1:1 beam splitter[22]. We extend the description of this phenomenon for multiple-OAM states carrying input photons. When OAM states of input photon $|\Phi\rangle$ is sent to the beam splitter, transmitted term A and reflected term B can be described with the following forms:

$$|\Phi\rangle \xrightarrow{\text{1:1 beam splitter}} \frac{1}{\sqrt{2}}|\Phi\rangle_A + \frac{i}{\sqrt{2}}R|\Phi\rangle_B . \tag{6}$$

where R represents flipping of OAM state, for example, $R|+1\rangle = |-1\rangle$. As shown in Fig. 5a, when OAM states of input photons are $|\Phi\rangle, |\Psi\rangle$ on the two BS inputs, the output state $|\Phi'\rangle \otimes |\Psi'\rangle$ can be described with the following forms:

$$\begin{aligned} |\Phi'\rangle \otimes |\Psi'\rangle &= \left(\frac{1}{\sqrt{2}}|\Phi\rangle_A + \frac{i}{\sqrt{2}}R|\Phi\rangle_B\right) \otimes \left(\frac{i}{\sqrt{2}}R|\Psi\rangle_A + \frac{1}{\sqrt{2}}|\Psi\rangle_B\right) \\ &= \left(\frac{i}{2}|\Phi\rangle_A \otimes R|\Psi\rangle_A\right) + \left(\frac{1}{2}|\Phi\rangle_A \otimes |\Psi\rangle_B - \frac{1}{2}R|\Psi\rangle_A \otimes R|\Phi\rangle_B\right) + \left(\frac{i}{2}R|\Phi\rangle_B \otimes |\Psi\rangle_B\right). \end{aligned} \tag{7}$$

With K being the number of OAM used in the system, the input states $|\Phi\rangle, |\Psi\rangle$ can be set to

$$|\Phi\rangle = \frac{1}{\sqrt{K}}\sum_{k=1}^{K} e^{i\phi_k}|+k\rangle, \quad |\Psi\rangle = \frac{1}{\sqrt{K}}\sum_{k=1}^{K} e^{i\psi_k}|-k\rangle, \tag{8}$$

considering that the two photons have the same polarization, wavelength, and are synchronized on the beam splitter. Each term of the output state given by Eq. 7 is described by the following:

$$\begin{aligned} |\Phi\rangle_A \otimes R|\Psi\rangle_A &= \left(\sum_{k=1}^{K}\frac{1}{\sqrt{K}}e^{i\phi_k}|+k\rangle_A\right) \otimes \left(\sum_{k=1}^{K}\frac{1}{\sqrt{K}}e^{i\psi_k}|+k\rangle_A\right) \\ &= \sum_{k=1}^{K}\frac{1}{K}e^{i(\phi_k+\psi_k)}|+k\rangle_A \otimes |+k\rangle_A + \sum_{k_1<k_2}^{K}\frac{1}{K}\left(e^{i(\phi_{k_1}+\psi_{k_2})} + e^{i(\psi_{k_1}+\phi_{k_2})}\right)|+k_1\rangle_A \otimes |+k_2\rangle_A \end{aligned} \tag{9}$$

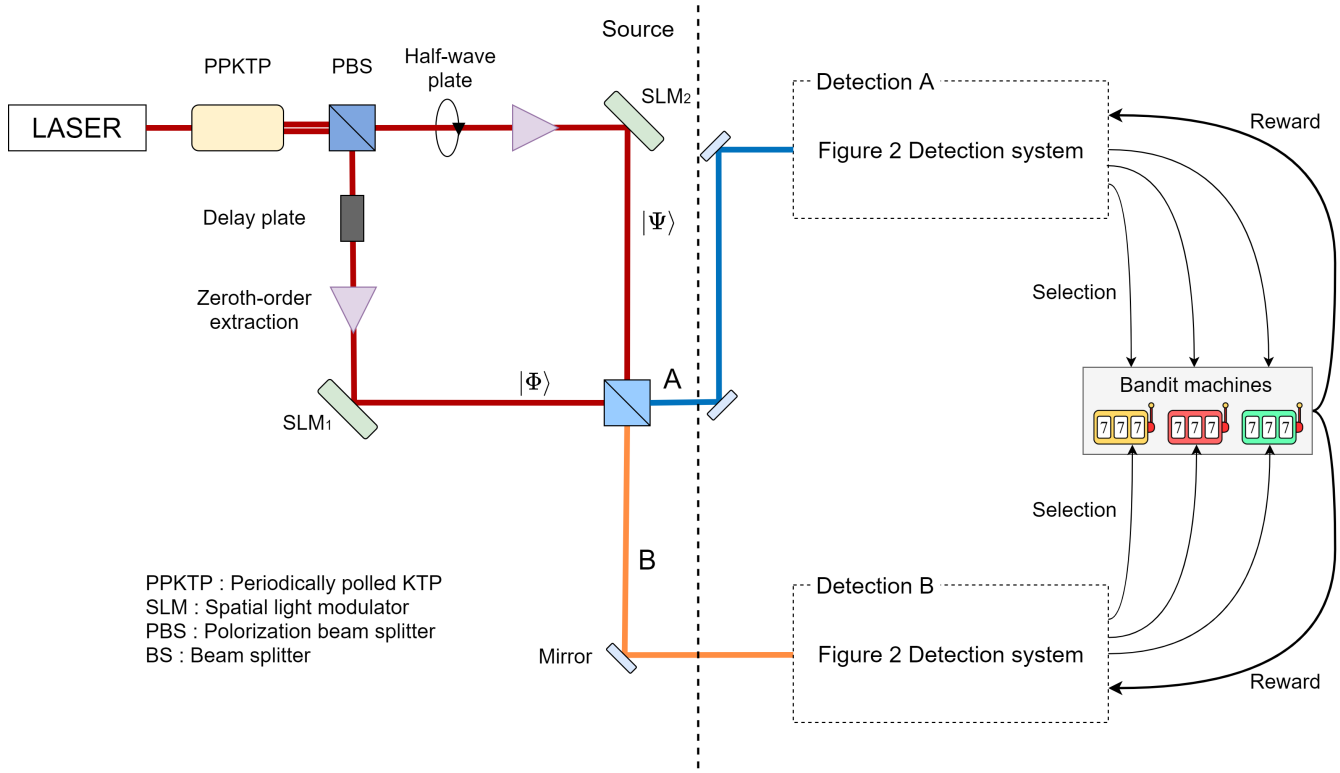
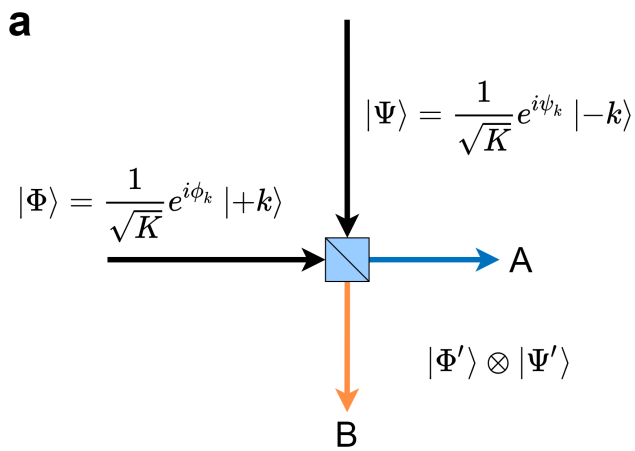


Figure 4. System architecture for solving 2-player 3-armed bandit problem.

$$
\begin{aligned}
|\Phi\rangle_A \otimes |\Psi\rangle_B - R|\Psi\rangle_A \otimes R|\Phi\rangle_B &= \sum_{k=1}^{K} \frac{1}{\sqrt{K}} e^{i\phi_k} |+k\rangle_A \otimes \sum_{k=1}^{K} \frac{1}{\sqrt{K}} e^{i\psi_k} |-k\rangle_B - \sum_{k=1}^{K} \frac{1}{\sqrt{K}} e^{i\psi_k} |+k\rangle_A \otimes \sum_{k=1}^{K} \frac{1}{\sqrt{K}} e^{i\phi_k} |-k\rangle_B \\
&= \sum_{k_1=1}^{K} \sum_{k_2=1}^{K} \frac{1}{K} e^{i(\phi_{k_1}+\psi_{k_2})} |+k_1\rangle_A \otimes |-k_2\rangle_B - \sum_{k_1=1}^{K} \sum_{k_2=1}^{K} \frac{1}{K} e^{i(\psi_{k_1}+\phi_{k_2})} |+k_1\rangle_A \otimes |-k_2\rangle_B \\
&= \sum_{k_1=1}^{K} \sum_{k_2=1}^{K} \frac{1}{K} \left(e^{i(\phi_{k_1}+\psi_{k_2})} - e^{i(\psi_{k_1}+\phi_{k_2})} \right) |+k_1\rangle_A \otimes |-k_2\rangle_B .
\end{aligned}
\tag{10}
$$

Therefore, the output state $|\Phi'\rangle \otimes |\Psi'\rangle$ is given by the following terms:

$$
\begin{aligned}
|\Phi'\rangle \otimes |\Psi'\rangle = &\sum_{k=1}^{K} \frac{i}{2K} e^{i(\phi_k+\psi_k)} |+k\rangle_A \otimes |+k\rangle_A \\
&+ \sum_{k_1<k_2}^{K} \frac{i}{2K} \left(e^{i(\phi_{k_1}+\psi_{k_2})} + e^{i(\psi_{k_1}+\phi_{k_2})} \right) |+k_1\rangle_A \otimes |+k_2\rangle_A \\
&+ \sum_{k_1=1}^{K} \sum_{k_2=1}^{K} \frac{1}{2K} \left(e^{i(\phi_{k_1}+\psi_{k_2})} - e^{i(\psi_{k_1}+\phi_{k_2})} \right) |+k_1\rangle_A \otimes |-k_2\rangle_B \\
&+ \sum_{k=1}^{K} \frac{i}{2K} e^{i(\phi_k+\psi_k)} |-k\rangle_B \otimes |-k\rangle_B \\
&+ \sum_{k_1<k_2}^{K} \frac{i}{2K} \left(e^{i(\phi_{k_1}+\psi_{k_2})} + e^{i(\psi_{k_1}+\phi_{k_2})} \right) |-k_1\rangle_B \otimes |-k_2\rangle_B .
\end{aligned}
\tag{11}
$$



b

	A	B	Detection probability
	$+k, +k$	—	$\frac{1}{2K^2}$
$(k_1 < k_2)$	$+k_1, +k_2$	—	$\frac{1}{K^2}\cos^2(\theta_{k_1} - \theta_{k_2})$
	$+k_1$	$-k_2$	$\frac{1}{K^2}\sin^2(\theta_{k_1} - \theta_{k_2})$
	—	$-k, -k$	$\frac{1}{2K^2}$
$(k_1 < k_2)$	—	$-k_1, -k_2$	$\frac{1}{K^2}\cos^2(\theta_{k_1} - \theta_{k_2})$

$$\left(\theta_k = \frac{\phi_k - \psi_k}{2}\right)$$

Figure 5. **(a)** Photons with OAM states $|\Phi\rangle, |\Psi\rangle$ go into BS as two inputs and go out to A and B. **(b)** Probability of detecting certain OAM states at each side of a beam splitter A and B. Some of the probabilities are constant, while the others depend on the phase difference between two input states. The probability of a pair of photons being detected at different sides is marked by the red frame.

Correspondingly, the probability of detecting the same state at the same side, that is $|+k\rangle_A \otimes |+k\rangle_A$ or $|-k\rangle_B \otimes |-k\rangle_B$, is given by

$$2 \cdot \left| \frac{i}{2K} e^{i(\phi_k + \psi_k)} \right|^2 = \frac{1}{2K^2}. \tag{12}$$

By introducing parameters $\theta_k = \frac{\phi_k - \psi_k}{2}$, which depends on the phase difference of two input states, the probability of detecting different states on the same side, that is $|+k_1\rangle_A \otimes |+k_2\rangle_A$ or $|-k_1\rangle_B \otimes |-k_2\rangle_B$, is given by

$$\left| \frac{i}{2K} \left(e^{i(\phi_{k_1} + \psi_{k_2})} + e^{i(\psi_{k_1} + \phi_{k_2})} \right) \right|^2 = \frac{1}{K^2} \cos^2(\theta_{k_1} - \theta_{k_2}), \tag{13}$$

and finally the probability of detecting pair of states on different sides, that is $|+k_1\rangle_A \otimes |-k_2\rangle_B$, is given by

$$\left| \frac{1}{2K} \left(e^{i(\phi_{k_1} + \psi_{k_2})} - e^{i(\psi_{k_1} + \phi_{k_2})} \right) \right|^2 = \frac{1}{K^2} \sin^2(\theta_{k_1} - \theta_{k_2}). \tag{14}$$

Figure 5b summarizes the probability of detecting each output state, while Fig. 6 shows all the probabilities with K ranging from 1 to 4. The probabilities depend only on θ_k, which can be tuned by controlling the SLM phases ϕ_k and ψ_k.

A pair of photons being detected on both sides is displayed with the red frames in Fig. 6, which are utilized as selections by the two players. What is remarkable is that the probability of detecting the same states at different sides is always zero because the probability term $\sin^2(\theta_k - \theta_k)$ is always equal to zero. For $K = 1$, this phenomenon corresponds to what is known as the Hong-Ou-Mandel effect. As the detected OAM states correspond to the selection of players, the probability of both players selecting the same machine is only limited by experimental constraints such as multiple pair generation, meaning that conflict-free decisions are accomplished.

The probabilities described in the red frames include the probabilities of detecting different states by the two players. It is remarkable that these probabilities can take equal value when K is less than or equal to three. For example, when $K = 2$, by assigning $\theta_1 = 0$ and $\theta_2 = \pi/2$, all such probabilities becomes $1/4$. Similarly, when $K = 3$, by setting $(\theta_1, \theta_2, \theta_3) = (0, \pi/3, 2\pi/3)$, the probabilities are all $1/12$. Namely, all arm combinations except selecting the same arm are selected equally. Note, however, that when K is larger or equal to four, we cannot perfectly equalize these probabilities by only tuning $\theta_1, \theta_2, \ldots, \theta_K$. This point is discussed in the *Discussion* section. In this study, we focus on the case when $K = 3$ because the equivalent selection of pairs is ensured, as discussed above.

$K = 1$

A	B	Detection probability
$+1, +1$	—	$\frac{1}{2}$
$+1$	-1	0
—	$-1, -1$	$\frac{1}{2}$

$K = 2$

A	B	Detection probability	When $\theta_k = \frac{k-1}{2}\pi$
$+1, +1$	—	$\frac{1}{8}$	$\frac{1}{8}$
$+1, +2$	—	$\frac{1}{4}\cos^2(\theta_1 - \theta_2)$	0
$+2, +2$	—	$\frac{1}{8}$	$\frac{1}{8}$
$+1$	-1	0	0
$+1$	-2	$\frac{1}{4}\sin^2(\theta_1 - \theta_2)$	$\frac{1}{4}$
$+2$	-1	$\frac{1}{4}\sin^2(\theta_1 - \theta_2)$	$\frac{1}{4}$
$+2$	-2	0	0
—	$-1, -1$	$\frac{1}{8}$	$\frac{1}{8}$
—	$-1, -2$	$\frac{1}{4}\cos^2(\theta_1 - \theta_2)$	0
—	$-2, -2$	$\frac{1}{8}$	$\frac{1}{8}$

$K = 3$

A	B	Detection probability	When $\theta_k = \frac{k-1}{3}\pi$
$+1, +1$	—	$\frac{1}{18}$	$\frac{1}{18}$
$+1, +2$	—	$\frac{1}{9}\cos^2(\theta_1 - \theta_2)$	$\frac{1}{36}$
$+1, +3$	—	$\frac{1}{9}\cos^2(\theta_1 - \theta_3)$	$\frac{1}{36}$
$+2, +2$	—	$\frac{1}{18}$	$\frac{1}{18}$
$+2, +3$	—	$\frac{1}{9}\cos^2(\theta_2 - \theta_3)$	$\frac{1}{36}$
$+3, +3$	—	$\frac{1}{18}$	$\frac{1}{18}$
$+1$	-1	0	0
$+1$	-2	$\frac{1}{9}\sin^2(\theta_1 - \theta_2)$	$\frac{1}{12}$
$+1$	-3	$\frac{1}{9}\sin^2(\theta_1 - \theta_3)$	$\frac{1}{12}$
$+2$	-1	$\frac{1}{9}\sin^2(\theta_1 - \theta_2)$	$\frac{1}{12}$
$+2$	-2	0	0
$+2$	-3	$\frac{1}{9}\sin^2(\theta_2 - \theta_3)$	$\frac{1}{12}$
$+3$	-1	$\frac{1}{9}\sin^2(\theta_1 - \theta_3)$	$\frac{1}{12}$
$+3$	-2	$\frac{1}{9}\sin^2(\theta_2 - \theta_3)$	$\frac{1}{12}$
$+3$	-3	0	0
—	$-1, -1$	$\frac{1}{18}$	$\frac{1}{18}$
—	$-1, -2$	$\frac{1}{9}\cos^2(\theta_1 - \theta_2)$	$\frac{1}{36}$
—	$-1, -3$	$\frac{1}{9}\cos^2(\theta_1 - \theta_3)$	$\frac{1}{36}$
—	$-2, -2$	$\frac{1}{18}$	$\frac{1}{18}$
—	$-2, -3$	$\frac{1}{9}\cos^2(\theta_2 - \theta_3)$	$\frac{1}{36}$
—	$-3, -3$	$\frac{1}{18}$	$\frac{1}{18}$

$K = 4$

A	B	Detection probability	When $\theta_k = \frac{k-1}{4}\pi$
$+1, +1$	—	$\frac{1}{32}$	$\frac{1}{32}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$+4, +4$	—	$\frac{1}{32}$	$\frac{1}{32}$
$+1$	-1	0	0
$+1$	-2	$\frac{1}{16}\sin^2(\theta_1 - \theta_2)$	$\frac{1}{32}$
$+1$	-3	$\frac{1}{16}\sin^2(\theta_1 - \theta_3)$	$\frac{1}{16}$
$+1$	-4	$\frac{1}{16}\sin^2(\theta_1 - \theta_4)$	$\frac{1}{32}$
$+2$	-1	$\frac{1}{16}\sin^2(\theta_1 - \theta_2)$	$\frac{1}{32}$
$+2$	-2	0	0
$+2$	-3	$\frac{1}{16}\sin^2(\theta_2 - \theta_3)$	$\frac{1}{32}$
$+2$	-4	$\frac{1}{16}\sin^2(\theta_2 - \theta_4)$	$\frac{1}{16}$
$+3$	-1	$\frac{1}{16}\sin^2(\theta_1 - \theta_3)$	$\frac{1}{16}$
$+3$	-2	$\frac{1}{16}\sin^2(\theta_2 - \theta_3)$	$\frac{1}{32}$
$+3$	-3	0	0
$+3$	-4	$\frac{1}{16}\sin^2(\theta_3 - \theta_4)$	$\frac{1}{32}$
$+4$	-1	$\frac{1}{16}\sin^2(\theta_1 - \theta_4)$	$\frac{1}{32}$
$+4$	-2	$\frac{1}{16}\sin^2(\theta_2 - \theta_4)$	$\frac{1}{16}$
$+4$	-3	$\frac{1}{16}\sin^2(\theta_3 - \theta_4)$	$\frac{1}{32}$
$+4$	-4	0	0
—	$-1, -1$	$\frac{1}{32}$	$\frac{1}{32}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
—	$-4, -4$	$\frac{1}{32}$	$\frac{1}{32}$

Figure 6. Probability of detecting certain OAM states at either sides of the beam splitter with K ranging from 1 to 4. The probability of a pair of photons being detected at different sides are marked by the red frame.

Simulation results for 2-player 3-armed bandit problem.

In the CMAB problem in the present study, the rewards are equally split among the players who selected the same machine; that is, the decision conflict by multiple players reduces the individual benefit. Furthermore, total rewards are reduced because of the conflicted choice.

Here we begin with a brief overview of the two-player decision-making situations by a game-theoretic formalism[42] while mentioning its intuitive implications. We denote $P_{k^*}, P_{k^{**}}, P_{k^{***}}$ respectively the first, second, and third highest reward probability. First, when $P_{k^*} > 2 \times P_{k^{**}}$, the situation of both players selecting machine 1 is the only Nash equilibrium. That is, conflict is unavoidable if both players act in a greedy manner because the best machine is far better than the other machines.

Second, when $P_{k^*} < 2 \times P_{k^{**}}$, Nash equilibrium is achieved when player 1 chooses the best machine (machine k^*), and player 2 selects the second-best machine (machine k^{**}), and vice versa. That is, conflicting decisions are avoided because changing the player's decision decreases his/her reward. However, there is a problem from the viewpoint of equality, as one of the players can keep selecting the higher reward machines while the other is locked with the lower reward decisions.

Third, there exists another symmetric Nash equilibrium with a mixed strategy, meaning that they select each machine with a certain probability. The details are described in the *Methods* section. Intuitively speaking, by this mixed strategy, both players sometimes intentionally refrain from choosing the best machine. Therefore, sometimes, decision conflicts can be avoided. Indeed, Lai *et al.* successfully utilized a mixed strategy in dynamic channel selection in communication systems[7]. However, it

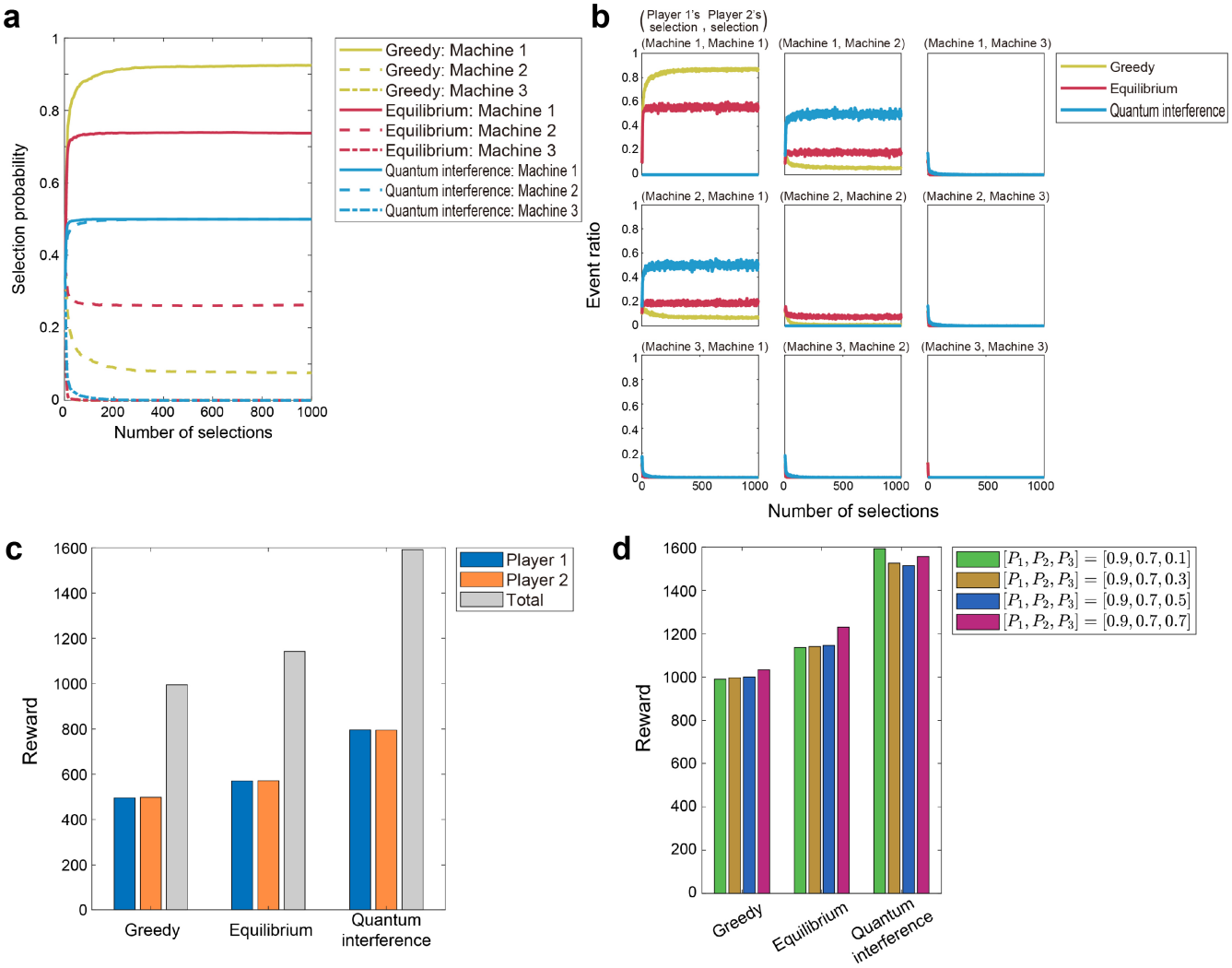


Figure 7. Simulation results for solving 2-player 3-armed competitive bandit problem with greedy, equilibrium and quantum interference policies. **(a)** Probability of selecting each machine over 1000 selections when $[P_1, P_2, P_3] = [0.9, 0.7, 0.1]$ averaged over 1000 repetitions. **(b)** Ratio of each selection pair over 1000 selections and 1000 repetitions. **(c)** Average rewards of player 1, player 2, and total rewards over 1000 selections with $[P_1, P_2, P_3] = [0.9, 0.7, 0.1]$ averaged over 1000 repetitions. **(d)** Average total rewards with different reward probability over 1000 selections and 5000 repetitions.

should be remarked that perfect conflict avoidance cannot be ensured by mixed strategies.

In order to quantitatively evaluate the performance differences among different policies, we compare the quantum interference system with the following two policies. One is a greedy policy where both players take greedy actions as if they are playing alone. The second is an equilibrium policy where both players try to achieve the symmetric Nash equilibrium by a mixed strategy. The details are described in the *Methods* section.

Figure 7 shows the results for solving the 2-player 3-armed bandit problem. Figure 7a shows how the selection probabilities of both players evolve with each policy. With the greedy policy, reminding that machine 1 has the highest reward probability of 0.9, its selection probability approaches almost 1 for both players, as in the case of a single player. For the equilibrium policy, the selection probabilities of the two most rewarding machines 1 and 2 converge to the probabilities defined by the mixed strategy. With the quantum interference strategy, however, machine 1 and machine 2 are selected with equal probability by both players. Figure 7b shows the ratio of each selection combination from both players. The greedy policy is associated with a large number of conflicts as both players almost only select machine 1, while the equilibrium policy reduces the number of conflicts to some extent as the selections are distributed. Finally, the quantum interference policy completely avoids conflicts. The final rewards with such selections are shown in Fig. 7c, for each player and for the total attributed reward. We observe that the quantum interference policy achieves almost ideal total rewards as well as equality between players. By contrast, the total

reward by the greedy and the equilibrium policies becomes small compared with the quantum interference policy because they suffer from unavoidable decision conflicts.

Figure 7d shows how the final reward of each policy varies when the reward probabilities of the three machines are modified. In greedy and equilibrium policies, the total reward changes due to the rate of selection of the lowest rewarding machine 3 in the exploration phase. On the other hand, with the quantum interference policy, the larger the difference between the reward probabilities of machine 2 and machine 3, the easier it is to determine the top two machines, and the higher the final total reward; despite this, the difference in total reward is mild in comparison with the difference with the other two policies.

Discussion

In this study, we show that we can benefit from the high dimensionality of OAM for scalability in solving multi-armed bandit problems. Furthermore, appropriate quantum interference constructions lead to achieving high rewards while maintaining a fair repartition between two players in competitive bandit problem situations. The total reward optimization is guaranteed by the selections of the two best machines by the two players in a non-conflicting manner, while the fair repartition is guaranteed by the equal probabilities of selection among players through quantum interference.

The main assumption is the simultaneous detection of exactly one photon for all players. In the proposed optical design, this is for the purpose of the extended Hong-Ou-Mandel effect or quantum interference that guarantees that identical photons go to the same side of the beam splitter, at the price of a post-selection of half of all photon pairs. While this is a strong constraint for potential applications, this design is only an example, and nothing forbids the obtention of the target state with other designs that do not rely on post-selection.

Regarding the extension to more arms, the current design is limited to three arms due to fundamental constraints (lack of enough degrees of freedom to constrain the 2-photon state). This may be solved by allowing to tune the relative amplitude between each OAM with the SLM and/or additional mechanisms. Once again, the goal of the setup presented in this study is only to present the principle of utilizing OAM for multiple arms in MAB and quantum interference for competitive decision-making. We believe that the extension to many arms is a technological problem without theoretical constraints[34].

The next discussion point is about security. The two-player CMAB system herein intends to let the players directly influence the detection probability via the attenuation amplitude in front of the detectors. While this architecture ensures independent machine selection and revision of the attenuation among the players, it presents one fundamental weakness: if a player only wants to select the highest rewarding arm, then the attenuation will be maximized for the lower arms, only letting photons reach the corresponding detector. However, this situation is easily identifiable by the other player, who can recognize that the probability of selecting a particular machine decreases. The solution for that player is straightforward: attenuate more the second-best arm too to correct the imbalance (in case of slight inequality), which is equivalent to not playing anymore if the other player completely blocks the other photons.

This brings the following discussion point about the photon utilization efficiency. In this study, only simultaneous detection of exactly one photon for all detectors of each and every player triggers the selection of arms from both players. The reason is to implement the post-selection of output states where one photon goes for both players instead of two photons for only one player. With the current operation principle based on quantum interference summarized in Fig. 6, half of all photon pairs are strictly unusable for the players. Although such a loss is unavoidable, further photon losses are induced in the system architecture shown in Fig. 4 because of the multiple BSs. As discussed earlier, this part can be improved by technological methods developed in the literature[33],[34].

Conclusion

To overcome the scalability limitations in the former single-photon-based decision making that relies on two orthogonal polarizations to resolve the two-armed bandit problem, we associate orbital angular momentum of photons to individual arms, which theoretically allows ideal scalability. When multiple players are involved, conflict of decisions becomes a serious issue, which is known as the competitive multi-armed bandit problem. Formerly, polarization-entangled photons have been shown to realize conflict-free decision making in two-player, two-armed situations; however, its arm-scalability is limited to only two. In this study, by extending the Hong-Ou-Mandel effect to more than two states, we theoretically establish an experimental configuration able to generate quantum interference among states with orbital angular momentum and conditions that provide conflict-free selections. We numerically examine total rewards regarding two-player, three-armed bandit problems, for which the proposed principle accomplishes almost the theoretical maximum, which is greater than a conventional mixed strategy intending to realize Nash equilibrium. This study paves a way toward photon-based intelligent systems as well as extending the utility of the high dimensionality of orbital angular momentum of photons and quantum interference in artificial intelligence domains.

Methods

Detail algorithm of greedy policy, equilibrium policy, and entanglement policy.

Greedy policy (Strategy for single player MAB).

Both players independently decide the probability of selecting each machine at each round. The algorithm is based on the softmax policy[5] and the probability of selecting machine i at round t is given by the following equation:

$$s_k(t+1) = \frac{e^{\beta \hat{P}_k(t)}}{\sum_{n=1}^{K} e^{\beta \hat{P}_n(t)}}. \tag{15}$$

If there exists a non-selected machine, all the machines are selected randomly with the same probability. Otherwise, $\hat{P}_k(t) = \frac{w_k(t)}{w_k(t)+l_k(t)}$ when the player has been rewarded $w_k(t)$ times and not rewarded $l_k(t)$ times from machine k. As a moderate tuning, β is set to be 20 in this study.

Equilibrium policy.

Expected reward for (Player 1, Player 2)		Selection of Player 2		
		Machine 1	Machine 2	Machine 3
Selection of Player 1	Machine 1	$(P_1/2, P_1/2)$	(P_1, P_2)	(P_1, P_3)
	Machine 2	(P_2, P_1)	$(P_2/2, P_2/2)$	(P_2, P_3)
	Machine 3	(P_3, P_1)	(P_3, P_2)	$(P_3/2, P_3/2)$

Table 1. Profit table of expected reward. Reward is splitted when the selection is in conflict between two players.

Table1 represents the profit table of the expected reward in a single selection. In Nash equilibrium, no player has anything to gain by changing only their own strategy. In Nash equilibrium, the strategy may be selecting one particular machine, but it could also be a strategy such that multiple machines are probabilistically chosen. In the situation of Table 1, strategies can be defined with the probabilities of selecting machine 1, machine 2, machine 3, or $\alpha_1, \alpha_2, \alpha_3$ for player A and $\beta_1, \beta_2, \beta_3$ for player B. In what follows, the notations k^*, k^{**}, k^{***} represent indices of the first, the second, and the third best machine, respectively. Nash equilibriums are summarized as shown below:

- in case $P_{k^*} > 2P_{k^{**}}$
  - ◇ $(\alpha_{k^*}, \alpha_{k^{**}}, \alpha_{k^{***}}) = (\beta_{k^*}, \beta_{k^{**}}, \beta_{k^{***}}) = (1, 0, 0)$
- in case $P_{k^*} < 2P_{k^{**}}$ and $P_{k^*}P_{k^{**}}/Q > \frac{2}{5}$
  - $(\alpha_{k^*}, \alpha_{k^{**}}, \alpha_{k^{***}}) = (1, 0, 0),\ (\beta_{k^*}, \beta_{k^{**}}, \beta_{k^{***}}) = (0, 1, 0)$
  - $(\alpha_{k^*}, \alpha_{k^{**}}, \alpha_{k^{***}}) = (0, 1, 0),\ (\beta_{k^*}, \beta_{k^{**}}, \beta_{k^{***}}) = (1, 0, 0)$
  - ◇ $(\alpha_{k^*}, \alpha_{k^{**}}, \alpha_{k^{***}}) = (\beta_{k^*}, \beta_{k^{**}}, \beta_{k^{***}}) = (\frac{2P_{k^*} - P_{k^{**}}}{P_{k^*} + P_{k^{**}}}, \frac{2P_{k^{**}} - P_{k^*}}{P_{k^*} + P_{k^{**}}}, 0)$
- in case $P_{k^*} < 2P_{k^{**}}$ and $P_{k^*}P_{k^{**}}/Q < \frac{2}{5}$
  - $(\alpha_{k^*}, \alpha_{k^{**}}, \alpha_{k^{***}}) = (1, 0, 0),\ (\beta_{k^*}, \beta_{k^{**}}, \beta_{k^{***}}) = (0, 1, 0)$
  - $(\alpha_{k^*}, \alpha_{k^{**}}, \alpha_{k^{***}}) = (0, 1, 0),\ (\beta_{k^*}, \beta_{k^{**}}, \beta_{k^{***}}) = (1, 0, 0)$
  - ◇ $(\alpha_{k^*}, \alpha_{k^{**}}, \alpha_{k^{***}}) = (\beta_{k^*}, \beta_{k^{**}}, \beta_{k^{***}}) = (2 - \frac{5P_{k^{**}}P_{k^{***}}}{Q}, 2 - \frac{5P_{k^{***}}P_{k^*}}{Q}, 2 - \frac{5P_{k^*}P_{k^{**}}}{Q})$

where $Q = P_{k^*}P_{k^{**}} + P_{k^{**}}P_{k^{***}} + P_{k^{***}}P_{k^*}$. With the equilibrium policy, both players try to achieve symmetric Nash equilibrium, which is represented with the shape ◇ above, under the situation that reward probabilities are not quite sure. In the simulation algorithm, each player decides which machines are better and which Nash equilibrium to achieve based on their own maximum likelihood estimation of reward probabilities. In an actual algorithm, the parameters of player 1 are calculated as below with $\hat{k}^*, \hat{k}^{**}, \hat{k}^{***}$ respectively representing machine indices with the first, the second, and the third highest estimated reward probability:

- in case $\hat{P}_{\hat{k}^*} > 2\hat{P}_{\hat{k}^{**}}$
  - $(\alpha^*, \alpha^{**}, \alpha^{***}) = (1, 0, 0)$
- in case $\hat{P}_{\hat{k}^*} > 2\hat{P}_{\hat{k}^{**}}$ and $\hat{P}_{\hat{k}^*}\hat{P}_{\hat{k}^{**}}/Q > \frac{2}{5}$
  - $(\alpha^*, \alpha^{**}, \alpha^{***}) = (\frac{2\hat{P}_{\hat{k}^*} - \hat{P}_{\hat{k}^{**}}}{\hat{P}_{\hat{k}^*} + \hat{P}_{\hat{k}^{**}}}, \frac{2\hat{P}_{\hat{k}^{**}} - \hat{P}_{\hat{k}^*}}{\hat{P}_{\hat{k}^*} + \hat{P}_{\hat{k}^{**}}}, 0)$
- in case $\hat{P}_{\hat{k}^*} < 2\hat{P}_{\hat{k}^{**}}$ and $\hat{P}_{\hat{k}^*}\hat{P}_{\hat{k}^{**}}/Q < \frac{2}{5}$
  - $(\alpha^*, \alpha^{**}, \alpha^{***}) = (2 - \frac{5\hat{P}_{\hat{k}^{**}}\hat{P}_{\hat{k}^{***}}}{Q}, 2 - \frac{5\hat{P}_{\hat{k}^{***}}\hat{P}_{\hat{k}^*}}{Q}, 2 - \frac{5\hat{P}_{\hat{k}^*}\hat{P}_{\hat{k}^{**}}}{Q})$

where $Q = \hat{P}_{\hat{k}^*}\hat{P}_{\hat{k}^{**}} + \hat{P}_{\hat{k}^{**}}\hat{P}_{\hat{k}^{***}} + \hat{P}_{\hat{k}^{***}}\hat{P}_{\hat{k}^*}$, and $\hat{P}_{\hat{k}^*}, \hat{P}_{\hat{k}^{**}}, \hat{P}_{\hat{k}^{***}}$ represent the first, the second, and the third highest estimated reward probability. The parameters of player 2 are also calculated in the same way with the different reward probability estimations. The probability of selecting each machine is calculated as below:

$$\begin{aligned}
s_{\hat{k}^*}(t+1) &= \alpha^*\pi(P_{\hat{k}^*} = P_{k^*}|H(t)) + \alpha^{**}\pi(P_{\hat{k}^*} = P_{k^{**}}|H(t)) + \alpha^{***}\pi(P_{\hat{k}^*} = P_{k^{***}}|H(t)) \\
s_{\hat{k}^{**}}(t+1) &= \alpha^*\pi(P_{\hat{k}^{**}} = P_{k^*}|H(t)) + \alpha^{**}\pi(P_{\hat{k}^{**}} = P_{k^{**}}|H(t)) + \alpha^{***}\pi(P_{\hat{k}^{**}} = P_{k^{***}}|H(t)) \\
s_{\hat{k}^{***}}(t+1) &= \alpha^*\pi(P_{\hat{k}^{***}} = P_{k^*}|H(t)) + \alpha^{**}\pi(P_{\hat{k}^{***}} = P_k{}^{**}|H(t)) + \alpha^{***}\pi(P_{\hat{k}^{***}} = P_{k^{***}}|H(t))
\end{aligned} \tag{16}$$

where, $\pi(P_{\hat{k}^*} = P_{k^*}|H(t))$ represents the probability of machine $\hat{k}^*$ to have the highest reward probability from the estimation based on the softmax policy. Here, $\pi(P_a > P_b > P_c|H(t))$ represents the probability of reward probabilities being $P_a > P_b > P_c$ under this estimation and it is calculated as below:

$$\pi(P_a > P_b > P_c|H(t)) = \frac{e^{\beta\hat{P}_a}}{e^{\beta\hat{P}_a} + e^{\beta\hat{P}_b} + e^{\beta\hat{P}_c}} \cdot \frac{e^{\beta\hat{P}_b}}{e^{\beta\hat{P}_b} + e^{\beta\hat{P}_c}}. \tag{17}$$

Therefore, probabilities of P_a being the first, the second, and the third best machine under estimation with softmax policy are:

$$\begin{aligned}
\pi(P_a = P_{k^*}|H(t)) &= \pi(P_a > P_b > P_c|H(t)) + \pi(P_a > P_c > P_b|H(t)) \\
\pi(P_a = P_{k^{**}}|H(t)) &= \pi(P_b > P_a > P_c|H(t)) + \pi(P_c > P_a > P_b|H(t)) \\
\pi(P_a = P_{k^{***}}|H(t)) &= \pi(P_b > P_c > P_a|H(t)) + \pi(P_c > P_b > P_a|H(t)).
\end{aligned} \tag{18}$$

Quantum interference policy.

In the quantum interference policy, both players try to select both the first and the second-best machine with the same probability to achieve fairness between the two players. Therefore, the probabilities of selecting machines are given by the following equation:

$$s_k(t+1) = \frac{1}{2}\left(\pi(P_k = P_{k^*}|H(t)) + \pi(P_k = P_{k^{**}}|H(t))\right). \tag{19}$$

References

1. Kitayama, K. *et al.* Novel frontier of photonics for data processing—photonic accelerator. *APL Photonics* **4**, 090901 (2019).
2. De Lima, T. F. *et al.* Machine learning with neuromorphic photonics. *J. Light. Technol.* **37**, 1515–1534 (2019).
3. Shen, Y. *et al.* Deep learning with coherent nanophotonic circuits. *Nat. Photonics* **11**, 441 (2017).
4. Van der Sande, G., Brunner, D. & Soriano, M. C. Advances in photonic reservoir computing. *Nanophotonics* **6**, 561–576 (2017).
5. Sutton, R. S. & Barto, A. G. *Introduction to Reinforcement Learning* (MIT Press, Cambridge, MA, USA, 1998), 1st edn.
6. Auer, P., Cesa-Bianchi, N. & Fischer, P. Finite-time analysis of the multiarmed bandit problem. *Mach. learning* **47**, 235–256 (2002).

7. Lai, L., El Gamal, H., Jiang, H. & Poor, H. V. Cognitive medium access: Exploration, exploitation, and competition. *IEEE Trans. Mob. Comput.* **10**, 239–253 (2010).

8. Kim, S.-J., Naruse, M. & Aono, M. Harnessing the computational power of fluids for optimization of collective decision making. *Philosophies* **1**, 245–260 (2016).

9. Naruse, M. *et al.* Single-photon decision maker. *Sci. Rep.* **5**, 13253 (2015).

10. Chauvet, N. *et al.* Entangled-photon decision maker. *Sci. Rep.* **9**, 12229 (2019).

11. Chauvet, N. *et al.* Entangled n-photon states for fair and optimal social decision making. *Sci. Rep.* **10**, 20420 (2020).

12. Naruse, M. *et al.* Single photon in hierarchical architecture for physical decision making: photon intelligence. *ACS Photonics* **3**, 2505–2514 (2016).

13. Allen, L., Barnett, S. M. & Padgett, M. J. *Optical angular momentum* (CRC press, 2003).

14. Forbes, A., de Oliveira, M. & Dennis, M. R. Structured light. *Nat. Photonics* **15**, 253–262 (2021).

15. Yao, A. M. & Padgett, M. J. Orbital angular momentum: origins, behavior and applications. *Adv. Opt. Photonics* **3**, 161–204 (2011).

16. Flamini, F., Spagnolo, N. & Sciarrino, F. Photonic quantum information processing: a review. *Rep. Prog. Phys.* **82**, 016001 (2019).

17. Forbes, A. & Nape, I. Quantum mechanics with patterns of light: progress in high dimensional and multidimensional entanglement with structured light. *AVS Quantum Sci.* **1**, 011701 (2019).

18. Krenn, M., Malik, M., Erhard, M. & Zeilinger, A. Orbital angular momentum of photons and the entanglement of laguerre–gaussian modes. *Phil. Trans. R. Soc. A* **375**, 20150442 (2017).

19. Zhang, Y. *et al.* Engineering two-photon high-dimensional states through quantum interference. *Sci. Adv.* **2**, e1501165 (2016).

20. Mirhosseini, M. *et al.* High-dimensional quantum cryptography with twisted light. *New J. Phys.* **17**, 033033 (2015).

21. Molina-Terriza, G., Torres, J. P. & Torner, L. Twisted photons. *Nat. Phys.* **3**, 305–310 (2007).

22. Hong, C.-K., Ou, Z.-Y. & Mandel, L. Measurement of subpicosecond time intervals between two photons by interference. *Phys. Rev. Lett.* **59**, 2044 (1987).

23. Pinheiro, A. *et al.* Vector vortex implementation of a quantum game. *JOSA B* **30**, 3210–3214 (2013).

24. Balthazar, W., Passos, M., Schmidt, A., Caetano, D. & Huguenin, J. Experimental realization of the quantum duel game using linear optical circuits. *J. Phys. B: At. Mol. Opt. Phys.* **48**, 165505 (2015).

25. Allen, L., Beijersbergen, M. W., Spreeuw, R. & Woerdman, J. Orbital angular momentum of light and the transformation of laguerre-gaussian laser modes. *Phys. Rev. A* **45**, 8185 (1992).

26. Beijersbergen, M., Coerwinkel, R., Kristensen, M. & Woerdman, J. Helical-wavefront laser beams produced with a spiral phaseplate. *Opt. Commun.* **112**, 321–327 (1994).

27. Heckenberg, N., McDuff, R., Smith, C. & White, A. Generation of optical phase singularities by computer-generated holograms. *Opt. Lett.* **17**, 221–223 (1992).

28. Beijersbergen, M. W., Allen, L., Van der Veen, H. & Woerdman, J. Astigmatic laser mode converters and transfer of orbital angular momentum. *Opt. Commun.* **96**, 123–132 (1993).

29. Padgett, M., Arlt, J., Simpson, N. & Allen, L. An experiment to observe the intensity and phase structure of laguerre–gaussian laser modes. *Am. J. Phys.* **64**, 77–82 (1996).

30. Wang, J. *et al.* Terabit free-space data transmission employing orbital angular momentum multiplexing. *Nat. Photonics* **6**, 488–496 (2012).

31. Leach, J., Padgett, M. J., Barnett, S. M., Franke-Arnold, S. & Courtial, J. Measuring the orbital angular momentum of a single photon. *Phys. Rev. Lett.* **88**, 257901 (2002).

32. Vaziri, A., Pan, J.-W., Jennewein, T., Weihs, G. & Zeilinger, A. Concentration of higher dimensional entanglement: qutrits of photon orbital angular momentum. *Phys. Rev. Lett.* **91**, 227902 (2003).

33. Lavery, M. P. *et al.* Refractive elements for the measurement of the orbital angular momentum of a single photon. *Opt. Express* **20**, 2110–2115 (2012).

34. Lavery, M. P. *et al.* Efficient measurement of an optical orbital-angular-momentum spectrum comprising more than 50 states. *New J. Phys.* **15**, 013024 (2013).

35. Cohen, J. D., McClure, S. M. & Yu, A. J. Should i stay or should i go? how the human brain manages the trade-off between exploitation and exploration. *Philos. Transactions Royal Soc. B: Biol. Sci.* **362**, 933–942 (2007).

36. Daw, N. D., O'doherty, J. P., Dayan, P., Seymour, B. & Dolan, R. J. Cortical substrates for exploratory decisions in humans. *Nature* **441**, 876–879 (2006).

37. Cesa-Bianchi, N., Gentile, C., Lugosi, G. & Neu, G. Boltzmann exploration done right. *arXiv preprint arXiv:1705.10257* (2017).

38. Pinnell, J., Rodríguez-Fajardo, V. & Forbes, A. Single-step shaping of the orbital angular momentum spectrum of light. *Opt. Express* **27**, 28009–28021 (2019).

39. Mair, A., Vaziri, A., Weihs, G. & Zeilinger, A. Entanglement of the orbital angular momentum states of photons. *Nature* **412**, 313–316 (2001).

40. Franke-Arnold, S., Barnett, S. M., Padgett, M. J. & Allen, L. Two-photon entanglement of orbital angular momentum states. *Phys. Rev. A* **65**, 033823 (2002).

41. Vaziri, A., Weihs, G. & Zeilinger, A. Experimental two-photon, three-dimensional entanglement for quantum communication. *Phys. Rev. Lett.* **89**, 240401 (2002).

42. Nash, J. F. *et al.* Equilibrium points in n-person games. *Proc. Natl. Acad. Sci.* **36**, 48–49 (1950).

Acknowledgements

This work was supported in part by the CREST Project (JPMJCR17N2) funded by the Japan Science and Technology Agency and Grants-in-Aid for Scientific Research (JP20H00233) funded by the Japan Society for the Promotion of Science.

Author contributions statement

M.N., N.C., and G.B. directed the project. T.A., N.C., G.B., S.H., and M.N designed the system architecture. T.A. and N.C conducted physical modeling and numerical performance evaluations. N.C., G.B., S.H., and R.H. examined technological constraints. All authors discussed the results. T.A., N.C., and M.N. wrote the manuscript. All authors reviewed the manuscript.

Additional information

Correspondence and requests for materials should be addressed to T.A. and M.N.
Competing interests: The authors declare no competing interests.