Quantum Speedup for Active Learning Agents

Giuseppe Davide Paparo,[1] Vedran Dunjko,[2,3,4] Adi Makmal,[2,3] Miguel Angel Martin-Delgado,[1] and Hans J. Briegel[2,3]

[1]*Departamento de Fisica Teorica I, Universidad Complutense, 28040 Madrid, Spain*
[2]*Institut für Theoretische Physik, Universität Innsbruck, Technikerstraße 25, A-6020 Innsbruck, Austria*
[3]*Institut für Quantenoptik und Quanteninformation der Österreichischen Akademie der Wissenschaften, A-6020 Innsbruck, Austria*
[4]*Division of Molecular Biology, Ruđer Bošković Institute, Bijenička cesta 54, 10002 Zagreb, Croatia*

Can quantum mechanics help us build intelligent learning agents? A defining signature of intelligent behavior is the capacity to learn from experience. However, a major bottleneck for agents to learn in real-life situations is the size and complexity of the corresponding task environment. Even in a moderately realistic environment, it may simply take too long to rationally respond to a given situation. If the environment is impatient, allowing only a certain time for a response, an agent may then be unable to cope with the situation and to learn at all. Here, we show that quantum physics can help and provide a quadratic speedup for active learning as a genuine problem of artificial intelligence. This result will be particularly relevant for applications involving complex task environments.

I. INTRODUCTION

The levels of modern-day technology have, in many aspects, surpassed the predictions made in the mid-20th century, as is easily witnessed, for example, by the sheer computing power of the average "smart" mobile phone. Arguably, the most striking exception to this, apart from, perhaps, human space exploration, lies in the development of genuine artificial intelligence (AI), the challenge of which has initially been greatly underestimated. The unceasing setbacks in the general AI problem caused research to shift emphasis to the production of useful technology, a direction now called applied AI. That is, emphasis was placed to specific algorithmic AI tasks— modules, such as data clustering, pattern matching, binary classification, and similar—and reduced from the holistic task of designing an autonomous and intelligent agent.

The discovery that the laws of quantum physics can be employed for dramatically enhanced ways of information processing [1–6] has already had a positive influence on specific algorithmic tasks of applied AI [7–12]. However, to our knowledge, it has not been demonstrated so far that quantum physics can help in the complemental task of designing autonomous and learning agents. The latter task is studied in the fields of embodied cognitive sciences and robotics [13–18], which promote a behavior-based approach to intelligence and put a strong emphasis on the physical aspects of an agent. The approach to AI we adopt in this work is along the lines of the latter perspective. We are guided by a few basic principles, inspired by biological agents, which include autonomy (implying that the agent must learn in, and adapt to, unknown dynamic environments), embodiedness (implying that the agent is situated in, and actively interacts with, a physical environment), and homogeneity (meaning that all possible separate "cognitive units" arise as possible configurations of one, or a few, homogeneous underlying systems that are, in principle, capable of growth). An example of a model that one could consider homogenous, aside from the projective-simulation model [19] we will consider here, would, for example, be artificial neural networks. One may then envision that true AI will emerge by growth and the learning of an agent, rather than through deliberate design. In this paper, we show that in such an embodied framework of AI, provable advancements of a broad class of learning agents can be achieved when we take the full laws of quantum mechanics into account.

II. LEARNING AGENTS AND QUANTUM PHYSICS

How could quantum physics help design better agents? An embodied agent is always situated in an environment from which it receives a sensory input, that is, a percept (from some set of percepts $\mathcal{S} = \{s_1, s_2, \ldots\}$), and, based on the received percept, it produces an action from the possible set of actions $\mathcal{A} = \{a_1, a_2, \ldots\}$; see Fig. 1. The capacity to learn implies that the agent is at every instant of time in some internal state that can change based on previous sequences of percept-action events. That is, it has a memory
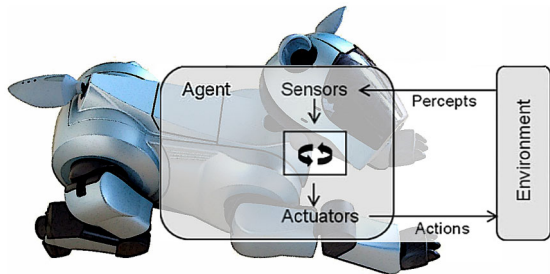
FIG. 1. An agent is always situated in an environment. It is equipped with sensors, through which it receives percepts from the environment, and with actuators, through which it can act on the environment. Based on perceptual input, the agent will after some internal processing engage its actuators and output an action. This figure is adapted and modified from Refs. [20,22].

that reflects the agent's history. The typical model for such autonomous learning agents we consider here is the reinforcement-learning model [20,21], where to each percept-action event, a reward in $\Lambda = \{0, 1\}$ (for simplicity, we consider binary rewards, but this reward system can be easily generalized in our model) is assigned when the action is correct.

Each percept-action-reward sequence constitutes an external time step (or cycle) of the activity of an agent. The learning process of the agent is characterized by an update rule of the internal state (based on the previous percept-action-reward sequences), and the (local) policy of the agent [21] is defined by what action is output, given the current internal state and the received percept. Unlike in typical reinforcement-learning settings, in embodied active agents, the required time to evaluate the policy (decide on an action) must be taken into account, and we refer to it as internal time.

The agent's learning process is reminiscent of computational oracle-query models, in which an unknown oracle (environment) is queried (via an action) by the agent, in an iterative quest for the best responses. It is tantalizing to consider employing the powerful quantum searching machinery [3,23,24], which has been proven to outperform classical algorithms in computational settings, in an attempt to improve the agent.

However, contrary to computer algorithms, an embodied agent, such as a robot, operates in a physical environment that is, for most existing applications, classical [25]. This prohibits querying in quantum superposition, a central ingredient to all quantum search algorithms. Thus, such naïve approaches to quantizing learning agents are doomed to fail [26].

Nonetheless, while the physical nature of the agent and the environment prohibits speedup through quantum queries of the environment, the physical processes within the agent, which lead to the performed actions, can be significantly improved by employing full quantum mechanics [27]. In particular, the required internal time

can be polynomially reduced in the model we present next. In general learning settings, this speedup alone will constitute an overall qualitative improvement of performance, for instance, when the environment changes on time scales not overwhelmingly larger than the agent's internal "thinking" time.

III. QUANTUM AGENTS BASED ON PROJECTIVE SIMULATION

A. The PS-agent model

The AI model of the agents we consider in the following is the so-called projective-simulation (PS) model [19], whose conceptual framework is in line with the desired guiding principles we highlighted earlier. The PS model is based on a specific memory system, which is called episodic and compositional memory (ECM). This memory provides the platform for simulating future action before real action is taken. The ECM can be described as a stochastic network of so-called clips that constitute the elementary excitations of episodic memory and can be implemented as excitations of suitable physical systems. The percepts ($\mathcal{S}$) and actions ($\mathcal{A}$), along with sequences thereof, are represented within an agent as the aforementioned clips, and the set of the clips comprises the clip space $C = \{c = (c^{(1)}, c^{(2)}, \ldots) | c^{(k)} \in \mathcal{S} \cup \mathcal{A}\}$. In this assumption work, we consider clips that are unit-length sequences, representing a memorized percept or an action, which we denote using the same symbols, so $C = \mathcal{S} \cup \mathcal{A}$ [28], but this can easily be generalized. The internal states of the agent, i.e., the total memory, comprise weighted graphs G_s over subsets of the clip space, which are assigned to each percept s. This graph dictates the hopping probabilities from one clip to another, and the hopping process realizes a Markov chain (MC). Thus, the elementary internal processes of the agent, which implement the transitions from clip to clip, are discrete-time stochastic diffusion processes. These diffusion processes can be realized in a variety of physical systems, as we discuss later.

The deliberation process of the agent is based only on the diffusion processes over the clip space of a certain (and, in general, variable) size, making this model homogeneous in the sense we explained earlier. The PS agents also perceive rewards $\Lambda = \{0, 1\}$ and, based on the perceived percept s, realized action a, and the resulting reward, the weights of the graph G_s (that is, transition probabilities) are updated via simple rules [29].

While the PS framework allows for many additional structures (see Appendix C for further details on the PS model), we will, for simplicity, only consider percept-specific flags—corresponding to rudimentary emoticons in Ref. [19]—which are subsets of actions assigned to each percept, formally $\mathcal{F} = \{f(s) | f(s) \subseteq \mathcal{A}, s \in \mathcal{S}\}$. Such flags may be used to represent the agent's short-term memory, in which case they significantly improve the performance of

the model [19]. For example, one can consider a possible mechanism in which, for each percept, all actions are initially flagged. If the agent outputs some action a, given a percept s, and this action is not rewarded, a is removed from $f(s)$. Once the set $f(s)$ has been depleted (indicating, for instance, that the environment has changed its policy), it is reset to contain all actions. The meaning of flags may be more general, and in this work, we only assume the sets of flags are always nonempty.

In the process of deliberation, the relevant Markov chain is diffused a particular number of times, depending on the particular PS model, until an action is output via so-called output couplers [30]. The choice of the action—and thereby the policy of the agent—is dictated by the probability distribution over the clip space, which is realized by the diffusion process. The latter depends on the agent's experience, manifest in the specified MC. Intuitively, this distribution represents the agent's state of belief regarding what is the right action in the given situation.

A particular model of PS we introduce here, so-called reflecting PS (RPS) agents, draws its name from the reflection process [19] in which the diffusion processes are repeated many times. Such agents approximate the complete mixing of their MCs, simulating infinite deliberation times. Once the mixing is (approximately) complete, the reflecting agent samples from the realized stationary distribution over the clip space (and, if needed, iterates the mixing process) until a flagged-action clip has been sampled. The internal states of reflecting PS agents are thus irreducible, aperiodic, and reversible MCs over subsets of the clip space (which contain all action clips). Reflecting PS agents can be seen as a generalization of so-called standard PS agents, and a comparison between a well-studied class of PS agents and RPS agents is provided in Appendix E.

In the limit of complete mixing, given a percept s and the current internal state (the MC P_s), the RPS agents output an action a distributed according to $\tilde{\pi}_s$ given as follows: Let π_s be the stationary distribution of P_s, and let $f(s)$ be the subset of flagged actions; then,

$$\tilde{\pi}_s(i) = \begin{cases} \dfrac{\pi_s(i)}{\sum_{j \in f(s)} \pi_s(j)} & \text{if } i \in f(s) \\ 0 & \text{otherwise,} \end{cases} \quad (1)$$

which is the renormalized stationary distribution π_s modified to only have support over flagged actions. We will often refer to $\tilde{\pi}_s$ as the tailed distribution.

In general, complete mixing is not possible or needed. To realize the (approximate) tailed distribution, as given in Eq. (1), the classical agent will iteratively have to prepare the approximate stationary distribution of P_s (by applying P_s to the initial distribution t^c_{mix} times) and sample from it, until a flagged action is hit. It is well known that, in order to mix the MC, t^c_{mix} should be chosen in $\tilde{O}(1/\delta_s)$ [31] (where

δ_s is the spectral gap of the MC P_s defined as $\delta_s = 1 - |\lambda_2|$ and λ_2 is the second-largest eigenvalue of P_s in absolute value), and the expected number of iterations of mixing and checking that have to be performed is $t^c_{\text{check}} \in \tilde{O}(1/\epsilon_s)$ [where $\epsilon_s = \sum_{i \in f(s)} \pi(i)$ is the probability of sampling a flagged action from the stationary distribution π_s]. Here, we will use the label P_s of the transition matrix as a synonym for the MC itself. Note that the internal time of the classical agent, i.e., the number of primitive processes (diffusion steps), is therefore governed by the quantities $1/\delta_s$ and $1/\epsilon_s$.

B. Quantum speedup of reflecting PS agents

The procedure the RPS agent performs in each deliberation step resembles a type of random-walk-based search algorithm that can be employed to find targeted items in directed weighted graphs [24]. In that context, the theory of quantum walks provides us with analogs of discrete-time diffusion processes, using which the search time can be quadratically reduced [23,24,32,33]. To design the quantum agent, inspired by these approaches to searching, here, we introduce a quantum-walk procedure that can be seen as a randomized Grover-like search algorithm (the quantum counterparts of classical search algorithms), most closely matching the main protocol in Ref. [24].

However, there are essential differences between searching problems and the problems of designing intelligent AI agents, which we expose and resolve in this work. First, we note that, for the task of simple searching, the procedure the RPS agent follows is known not to be optimal, in general [24,34]. In contrast, for the task of the RPS agent, which is to output a flagged action, according to a good approximation of the tailed distribution in Eq. (1), this algorithm is, in general, optimal [35]. The optimality can be seen by the known lower bounds for mixing times t^c_{mix} of reversible MCs. (See Appendix B for details.)

Furthermore, while as a direct consequence of the results in Ref. [24], the quantum RPS produces a flagged action in time that is quadratically faster than that achieved using the procedure employed by the classical agent, prior works provide no guarantees that the output actions will be distributed according to the desired tailed distribution. Recall, in the context of AI, that all agents produce some action, and it is precisely the output distribution that differentiates one agent from another in terms of behavior (and thus, success). In this work, we prove that both the output distributions of the reflecting classical and quantum agents approximate the distribution of Eq. (1) and thus are approximately equal. That is, they belong to the same behavioral class. (For a formal definition of the behavioral classes that we introduce, see Appendix D.) Consequently, the quantum reflecting-agent construction we give realizes, in a full sense, quantum-enhanced analogs of classical reflecting agents.

While quantum approaches to sampling problems have not until now been exploited in AI, we observe that the methodology we use is, in spirit, related to the problem of sampling from particular distributions. These approaches have been extensively explored, often in the context of Markov-chain Monte Carlo methods, where quantum speedup can also be obtained [36–39]. There, the quantum walks were mostly used for the important purpose of sampling from Boltzmann-Gibbs distributions of (classical and quantum) Hamiltonians.

In order to define the quantum reflecting agent, we first review the standard constructions and results from the theory of classical and quantum random walks [23,24,33] and refer the reader to Appendix B for more details.

The quantum RPS agent we propose uses the standard quantum discrete-time diffusion operators U_{P_s} and V_{P_s} that act on two quantum registers, sufficiently large enough to store the labels of the nodes of the MC P_s. The diffusion operators are defined as $U_{P_s}|i\rangle|0\rangle = |i\rangle|p_i\rangle$ and $V_{P_s}|0\rangle|j\rangle = |p_j^*\rangle|j\rangle$, where $|p_i\rangle = \sum_j \sqrt{[P_s]_{ji}}|j\rangle$ and $|p_j^*\rangle = \sum_i \sqrt{[P_s^*]_{ij}}|i\rangle$. Here, P_s^* is the time-reversed MC defined by $\pi_i P_{ji} = \pi_j P_{ij}^*$, where $\pi = (\pi_i)_i$ is the stationary distribution of P_s [40]. Using four applications of the diffusion operators above, it has been shown that one can construct the standard quantum-walk operator $W(P_s)$ (sometimes referred to as the Szegedy-walk operator), which is a composition of two reflections in the mentioned two-register state space. In particular, let Π_1 be a projection operator on the space $\text{span}\{|i\rangle|p_i\rangle\}_i$ and Π_2 be the projector on $\text{span}\{|p_j^*\rangle|j\rangle\}_j$. Then, $W(P_s) = (2\Pi_2 - \mathbb{1}) \times (2\Pi_1 - \mathbb{1})$.

Using the quantum-walk operator and the well-known phase-detection algorithm [5,24], the agent can realize the $R(P_s)(q,k)$ subroutine that approximates the reflection operator $2|\pi_s\rangle\langle\pi_s| - \mathbb{1}$, where $|\pi_s\rangle = \sum_i \sqrt{\pi_s(i)}|i\rangle$ is the coherent encoding of the stationary distribution π_s. The parameters q and k control the fidelity (and the time requirement) of this process, i.e., how well the reflection is approximated as a function of the number of applications of the quantum-walk operator.

More precisely, in the implementation of the $R(P_s)(q,k)$ operator, the quantum RPS agent utilizes an ancillary register of $q \times k$ qubits. To ensure the correct behavior, q is chosen as $q \in \tilde{O}(1/\sqrt{\delta_s})$, which depends on the square root of the spectral gap of the MC P_s. Under this condition, it has been shown that the distance between the ideal reflection operator and the realized approximate operator is upper bounded by 2^{1-k}, under a suitable metric. (See Appendix B, Theorem 3, for details.) That is, the fidelity of this reflection operator approaches unity exponentially quickly in the parameter k [24].

To produce a flagged action according to the desired distribution, the quantum RPS agent will first initialize its quantum register to the state $|\pi_{\text{init}}\rangle = U_{P_s}|\pi_s\rangle|0\rangle$, which requires just one application of the diffusion operator U_{P_s}, provided the state $|\pi_s\rangle = \sum_i \sqrt{\pi_s(i)}|i\rangle$ is available. Here, like in the standard frameworks of algorithms based on quantum walks with nonsymmetric Markov chains, we assume that the state $|\pi_s\rangle$ is available, and in Appendix E, we provide an example of reflecting classical and quantum agents where this is easily achieved [41].

Following the initial step, the agent performs a randomized Grover-like sequence of reflections, reflecting over flagged actions (denoted as $\text{ref}[f(s)]$), interlaced with reflections using the approximate reflection operator described previously. After the reflections have been completed the required number of times, the resulting state is measured and the found flagged action is output. In the case that a nonaction clip is hit, the entire procedure is repeated [42].

Since Grover-like search engines guarantee that the overlap between the final state and a state with support just over the actions is constant, the probability of not hitting a flagged action decreases exponentially quickly in the total number of iterations and does not contribute significantly to our analysis. In Appendix D, we provide a detailed analysis and propose a method for the efficient repreparation of the required initial state (by recycling of the residual state), in the event that the deliberation procedure should be repeated. The deliberation process is detailed in Fig. 2. The total number of reflections required is $t_{\text{check}}^q \in_R [0, \tilde{O}(1/\sqrt{\epsilon_s})]$ (this choice is uniform at random, as for the randomized Grover algorithm [43]), and the approximate reflection operator is applied with parameters $R(P_s)[t_{\text{mix}}^q, \tilde{O}(1)]$, where $t_{\text{mix}}^q \in \tilde{O}(1/\sqrt{\delta_s})$ [44].

Therefore, the total number of calls to the diffusion operators U_{P_s} and V_{P_s} is in $\tilde{O}(1/(\sqrt{\epsilon_s \delta_s})$ [and the total number of reflections over flagged actions—equivalents of checks in the classical agent—is in $\tilde{O}(1/\sqrt{\epsilon_s})$], which is a quadratic improvement over the classical agent. As we have previously mentioned, the remaining key ingredient to our result is the fact that the proposed quantum agent produces actions (approximately) according to the tailed distribution in Eq. (1) and that the approximations for both agents can
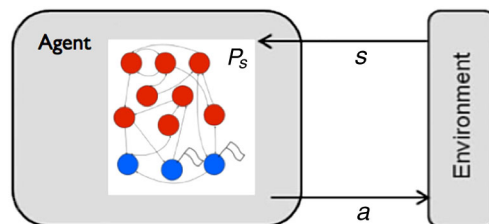

FIG. 2.   Both the classical and the quantum RPS are characterized by their internal Markov chains over the clip space. The procedures the quantum agent performs in order to choose the desired actions are listed. The values t_{mix}^q and t_{check}^q are chosen in $\tilde{O}(1/\sqrt{\delta_s})$ and uniformly at random in $[0, \tilde{O}(1/\sqrt{\epsilon_s})]$, respectively; see the text for details.

be made arbitrarily good within, at most, logarithmic overhead. We leave the proof of this claim for Appendix D. We note that in this paper, we have presented constructions for reversible MCs for simplicity, but the constructions can be extended to general irreducible chains using approaches analogous to those in Ref. [24].

We have thus presented a method for the generic quantization of reflecting PS agents that maintains the behavior of the agents and provably yields a quadratic speedup in internal times that is vital in real environment settings.

IV. DISCUSSION

We have presented a class of quantum learning agents that use quantum memory for their internal processing of previous experience. These agents are situated in a classical task environment that rewards a certain behavior but is otherwise unknown to the agent, which corresponds to the situation of conventional learning agents.

The agent's internal "program" is realized by physical processes that correspond to quantum walks. These quantum walks are derived from classical random walks over directed weighted graphs that represent the structure of its episodic memory. We have shown, using quantum coherence and known results from the study of quantum walks, how the agent can explore its episodic memory in superposition in a way that guarantees a provable quadratic speedup in its active learning time over its classical analog.

Regarding potential realizations for such quantum learning agents, modern quantum physics laboratories are exploring varieties of systems that can serve as suitable candidates. Quantum random walks and related processes can naturally be implemented in linear optics setups by, for instance, arrays of polarizing beam splitters [45], and highly versatile setups can also be realized using internal states of trapped ions [46]. Such advancements, all of which belong to the field of quantum simulation [47], could be used as ingredients toward the implementation of quantum reflecting agents, without the need to develop a full-blown universal quantum computer.

An entirely different route toward realizing the proposed quantum (and classical) learning agents might employ condensed-matter systems in which the proposed Markov chains could, e.g., be realized through cooling or relaxation processes toward target distributions that then encode the state of belief of the agent. Here, we envision rather nontrivial cooling or relaxation schemes in complex many-body systems, the study of which is also a prominent topic in the field of quantum simulation.

In conclusion, it seems to us that the embodied approach to artificial intelligence acquires a further fundamental perspective by combining it with concepts from the field of quantum physics. The implications of embodiment are, in the first place, described by the laws of physics, which tell us not only about the constraints but also the ultimate possibilities of physical agents. In this paper, we have shown an example of how the laws of quantum physics can be fruitfully employed in the design of future intelligent agents that will outperform their classical relatives in complex task environments.

APPENDIX A: FORMAL DEFINITIONS AND BEHAVIOR OF REINFORCEMENT-LEARNING AGENTS

Here, we formally define the model of reinforcement-learning agents as employed in this work.

Definition 1 (Reinforcement-learning agent).—A reinforcement-learning agent is an ordered sextuplet $(\mathcal{S}, \mathcal{A}, \Lambda, \mathcal{C}, \mathcal{D}, \mathcal{U})$, where (i) $\mathcal{S} = \{s_1, \ldots, s_m\}$ and $\mathcal{A} = \{a_1, \ldots, a_n\}$ are the sets of percepts and actions, respectively; (ii) $\Lambda = \{0, 1\}$ is the set of rewards, offered by the environment; (iii) $\mathcal{C} = \{C_1, \ldots, C_p\}$ is the set of possible internal states of the agent; (iv) $\mathcal{D} : \mathcal{S} \times \mathcal{C} \to \mathcal{A}$ is the decision function that outputs some action, given a percept and the internal state; and (v) $\mathcal{U} : \mathcal{S} \times \mathcal{A} \times \Lambda \times \mathcal{C} \to \mathcal{C}$ is the update function that updates the internal state based on the success or failure of the last percept-action sequence.

A few comments are in order. In this work, the sets of percepts, actions, and internal states are defined to be finite, but, in general, this need not be the case. The set of rewards is binary, and this choice can again be generalized. The update function may take additional information into account, based on additional outputs of the decision function, which are only processed internally, but this does not occur in the models we consider.

The decision function is not necessarily deterministic. In the nondeterministic case, it can be formally defined as

$$\mathcal{D} : \mathcal{S} \times \mathcal{C} \to \mathrm{distr}(\mathcal{A}), \tag{A1}$$

that is, a function that takes values in the set of distributions over $\mathcal{A}$. In this case, we also assume that this distribution is sampled before actual output is produced and that the sampled action is the input to the update function.

Next, we consider equivalences between agents in so-called passive settings.

In the algorithmic tradition of machine learning, the learning pace is measured by external time (steps) alone and the typical figure of merit is the percentage of rewarded actions of the agent, as the function of external time. From an embodied-agent perspective, this set-up corresponds to a special passive setting where a static environment always waits for the responses of the agent. This constraint imposes a restriction on the universality of statements that can at all be made about the performance of an agent. In particular, in that setting, it is well known that no two agents can be meaningfully compared without reference to a specific learning task (or a collection of them)—a collection of results dubbed "no free lunch theorems" and "almost no free lunch theorems" [48–50]. These results prove that when one agent outperforms another in a certain environment, there exists a different environment where the ranking according to performance is reversed. This fact, for instance, implies that every choice of environment settings, for which the results of agent performance are presented, must first be well justified. More critically for our agenda, in which we wish to make no assumptions on the environment, passive settings would imply that no comparative statements relating the performances of agents are possible. In active scenarios, internal time does matter, but, nonetheless, the passive setting plays a part. It provides a baseline for defining a passive behavioral equivalence of agents that will be instrumental in our analysis of active scenarios.

Let us denote the elapsed sequence of triplets (percept, action, reward) that had occurred up to the time step k (the history of the agent) with H_k, for two agents A and A' who can perceive and produce the same sets of percepts ($\mathcal{S}$) and actions ($\mathcal{A}$), respectively. Then, we will say that A and A' are passively ($\epsilon-$) equal if at every external time step k, the probabilities P_A and $P_{A'}$ of agents A and A', respectively, outputting some action $a \in \mathcal{A}$, given every percept $s \in \mathcal{S}$, and given all possible identical histories H_k, are ($\epsilon-$) equal, in terms of the variational distance on distributions:

$$\frac{1}{2}\sum_a |P_A(a|H_k,s) - P_{A'}(a|H_k,s)| \leq \epsilon, \qquad \text{(A2)}$$

which we abbreviate with

$$A \approx_\epsilon A'. \qquad \text{(A3)}$$

If the agents considered are equipped with an extra parameter τ (a precision parameter), which fine-tunes the behavior of the agent, we can demand more and require that the approximate equality above converge to an equality (i.e., $\epsilon \to 0$, as $\tau \to \infty$). Then, in the limit, the relation above induces passive behavioral equivalence classes for fixed sets of possible percepts and actions. In the case of the classical and quantum agents we consider in the main text, such precision parameters do exist, and, as we show later in

these Appendixes, the approximate equality converges to an equality.

In passive settings, by definition, two passively equal agents perform equally well, and comparison of agents within a class is pointless. However, in the active scenario, with the classes in place, we can meaningfully compare agents within the same class, with no assumptions on the environment [51]. Indeed, in an active learning setting, two passively equal agents A and A' may have vastly different success chances. To see this difference at work, suppose that the environment changes its policies on a time scale that is long compared to the internal time scale of agent A but short relative to the internally slower agent A'. The best policy of both agents is to query the environment as frequently as possible, in order to learn the best possible actions. However, from the perspective of the slow agent, the environment will look fully inconsistent—once-rewarded actions are no longer the right choice, as that agent simply did not have the time to learn. Thus, in active scenarios, the internal speed of the agent is vital.

APPENDIX B: CLASSICAL AND QUANTUM-WALK BASICS

A random walk on a graph is described by a MC, specified by a transition matrix P that has entries $P_{ji} = \text{prob}(j|i)$. For an irreducible MC, there exists a stationary distribution π such that $P\pi = \pi$. For an irreducible and aperiodic MC, this distribution can be approximated by $P^t\pi_0$, i.e., by applying, to any initial distribution π_0, the MC t number of times, where $t \geq t_{\epsilon'}^{\text{mix}}$. This time is known as the mixing time and is defined as follows.

Definition 2 (Mixing time).—The mixing time is

$$t_{\epsilon'}^{\text{mix}} = \min\{t|\forall s \geq t, \pi_0 : ||P^s\pi_0 - \pi|| \leq \epsilon'\}.$$

The latter can be related to the spectral properties of the MC P, in the case of reversible chains, via the following theorem [52].

Theorem 1.—The mixing time satisfies the following inequalities:

$$\frac{1}{\delta}\frac{\lambda_2}{\log 2\epsilon'} \leq t_{\epsilon'}^{\text{mix}} \leq \frac{1}{\delta}[\max \log \pi_i^{-1} + \log(\epsilon')^{-1}].$$

Here, we use ϵ' instead of the standard ϵ for consistency with the rest of the Appendixes, as ϵ has a reserved meaning.

For the purpose of clarity, let us introduce some definitions and theorems, originally provided in Refs. [23,24], that will be useful to introduce the notation and to prove the main results for the speedup of quantum agents.

The quantum analog of the application of the MC P is given by the following definition.

Definition 3 (Quantum diffusion operators).—The quantum diffusion operators, the analogs of the classical diffusion operators, are given by the following transformations:

$$U_P: \ |i\rangle|0\rangle \mapsto |i\rangle|p_i\rangle, \tag{B1}$$

$$V_P: \ |0\rangle|j\rangle \mapsto |p_j^*\rangle|j\rangle, \tag{B2}$$

where $|p_i\rangle = \sum_j \sqrt{P_{ji}}|j\rangle$, $|p_j^*\rangle = \sum_i \sqrt{P_{ij}^*}|i\rangle$, and P^* is the time-reversed MC defined by $\pi_i P_{ji} = \pi_j P_{ij}^*$.

We will consider the application of the MC P (for the classical agent) and the quantum diffusion operators (for the quantum agent) as the (equally time-consuming) primitive processes, as is done in the theory of quantum random walks [24]. Next, we can define the quantum-walk operator $W(P)$, for the Markov chain P.

Definition 4 (Walk operator or quantum Markov chain).—The walk operator or quantum Markov chain is given by

$$W(P) = (2\Pi_2 - \mathbb{1})(2\Pi_1 - \mathbb{1}), \tag{B3}$$

where Π_1 is the projection operator onto the space $\mathrm{span}\{|i\rangle|p_i\rangle\}_i$ and Π_2 is the projection operator onto $\mathrm{span}\{|p_j^*\rangle|j\rangle\}_j$.

The quantum-walk operator can be easily realized through four applications of the quantum diffusion operators; see, e.g., Ref. [24] for details.

Another standard operation that both the classical and the quantum agents perform is checking whether the clip found is flagged (corresponding to checking whether an item is marked). The quantum check operator is defined as follows.

Definition 5 (Check).—The quantum check operator is the reflection denoted as $\mathrm{ref}[f(s)]$ performing

$$|i\rangle|j\rangle \mapsto \begin{cases} -|i\rangle|j\rangle & \text{if } i \in f(s) \\ |i\rangle|j\rangle & \text{otherwise} \end{cases}, \tag{B4}$$

where $f(s)$ denotes the set of flagged actions corresponding to the percept s.

In order to prove our main theorems, we will be using the ideas introduced in the context of quantum searching [23,24], which we now briefly expose. In the quantum-walk-over-graphs approach to searching, one defines an initial state that encodes the stationary distribution of a MC

$$|\pi\rangle = \sum_i \sqrt{\pi(i)}\,|i\rangle|p_i\rangle \tag{B5}$$

and performs a rotation onto the state containing the "marked items"

$$|\tilde{\pi}\rangle := \frac{\Pi^{f(s)}|\pi\rangle}{\|\Pi^{f(s)}|\pi\rangle\|}, \tag{B6}$$

where $\Pi^{f(s)}$ is the projector on the space of marked items, i.e., $\Pi^{f(s)} = \sum_{i \in f(s)} |i\rangle\langle i| \otimes \mathbb{1}$. Let us point out that $\mathrm{span}\{|\tilde{\pi}\rangle, |\pi\rangle\} \subseteq \mathrm{span}\{|i\rangle|p_i\rangle\}_i + \mathrm{span}\{|p_j^*\rangle|j\rangle\}_j$.

In order to achieve this rotation, one makes use of two reflections. The first is the reflection over $|\tilde{\pi}^\perp\rangle$ [denoted as $\mathrm{ref}(|\tilde{\pi}^\perp\rangle)$], the state orthogonal to $|\tilde{\pi}\rangle$ in $\mathrm{span}\{|\tilde{\pi}\rangle, |\pi\rangle\}$. This operator can be realized using the primitive of checking. Indeed, we have the following claim (stated in Ref. [24]), given by the following lemma.

Lemma 1.—Restricted on the subspace $\mathrm{span}\{|\tilde{\pi}\rangle, |\pi\rangle\}$, the action of $\mathrm{ref}(|\tilde{\pi}^\perp\rangle)$ is identical to $-\mathrm{ref}[f(s)]$.

Proof.—Let $\alpha|\pi\rangle + \beta|\tilde{\pi}\rangle$ be a vector in $\mathrm{span}\{|\tilde{\pi}\rangle, |\pi\rangle\}$. We have that

$$-\mathrm{ref}[f(s)](\alpha|\pi\rangle + \beta|\tilde{\pi}\rangle) = \alpha(\mathbb{1} - 2\Pi^{f(s)})|\pi\rangle - \beta|\tilde{\pi}\rangle, \tag{B7}$$

where $\Pi^{f(s)}$ is the projector on the set $f(s)$. The result easily follows by noting that

$$\mathrm{ref}(|\tilde{\pi}^\perp\rangle)(\alpha|\pi\rangle + \beta|\tilde{\pi}\rangle)$$
$$= \alpha\left(\mathbb{1} - 2\frac{\Pi^{f(s)}|\pi\rangle\langle\pi|\Pi^{f(s)}}{\|\Pi^{f(s)}|\pi\rangle\|^2}\right)|\pi\rangle - \beta|\tilde{\pi}\rangle \tag{B8}$$

and that

$$\frac{\Pi^{f(s)}|\pi\rangle\langle\pi|\Pi^{f(s)}}{\|\Pi^{f(s)}|\pi\rangle\|^2}|\pi\rangle = \frac{\langle\pi|\Pi^{f(s)}|\pi\rangle}{\|\Pi^{f(s)}|\pi\rangle\|^2}\Pi^{f(s)}|\pi\rangle = \Pi^{f(s)}|\pi\rangle,$$

since $\frac{\langle\pi|\Pi^{f(s)}|\pi\rangle}{\|\Pi^{f(s)}|\pi\rangle\|^2} = 1$. ∎

On the other hand, the reflection over $|\pi\rangle$ is not straightforward. One can devise an approximated scheme to implement this reflection using the phase-estimation algorithm. Indeed, one can build a unitary operator, using phase estimation applied to the quantum-walk operators, which approximates the reflection over $|\pi\rangle$. Before we state the theorem regarding this approximate reflection operator (constructively proven in Ref. [24]), we will first give another result regarding the spectrum of the quantum-walk operator, which will be relevant to us presently.

Theorem 2 (Szegedy [23]).—Let P be an irreducible, reversible MC with stationary distribution π. Then, the quantum-walk operator $W(P)$ is such that (i) $W(P)|\pi\rangle = |\pi\rangle$; (ii) $W(P)|\psi\rangle = \exp(\pm 2i\theta)|\psi\rangle$, where $\cos(\theta) = |\lambda|$ is the absolute value of an eigenvalue of P and $|\psi\rangle \in \mathrm{span}\{|i\rangle|p_i\rangle\}_i + \mathrm{span}\{|p_j^*\rangle|j\rangle\}_j$; and (iii) $W(P)$ has no other eigenvalue in $\mathrm{span}\{|i\rangle|p_i\rangle\}_i + \mathrm{span}\{|p_j^*\rangle|j\rangle\}_j$.

Note that the phase gap Δ, defined as the minimum nonzero 2θ, is such that $\cos \Delta = |\lambda_2|$, where λ_2 is the second-largest eigenvalue of P with respect to the absolute

value. One can then, with some algebra, conclude that $\Delta \geq 2\sqrt{\delta}$.

Let us note that any unitary able to approximately detect whether the eigenvalue of $W(P)$ of a state in $\text{span}\{|i\rangle|p_i\rangle\}_i + \text{span}\{|p_j^*\rangle|j\rangle\}_j$ is different from 1 (or, equivalently, its eigenphase is different from 0) and conditionally flip the state will do. We will use such a unitary to approximate $\text{ref}(|\pi\rangle)$. Let us use this intuition to build such a unitary $R(P)$ that takes as a parameter the precision s and refer to it in the following as the approximate reflection operator $R(P)$ [24].

Theorem 3 (Approximate reflection operator [24]).—Let P be an ergodic, irreducible Markov chain on a space of size $|X|$ with (unique) stationary distribution π. Let $W(P)$ be the corresponding quantum Markov chain with phase gap Δ. Then, if s is chosen in $\mathcal{O}[\log_2(1/\Delta)]$, for every integer k, there exists a unitary $R(P)$ that acts on $2\lceil \log_2 |X| \rceil + ks$ qubits, such that (i) $R(P)$ makes at most $k2^{s+1}$ calls to the (controlled) $W(P)$ and $W(P)^\dagger$; (ii) $R(P)|\pi\rangle|0\rangle^{\otimes ks} = |\pi\rangle|0\rangle^{\otimes ks}$; and (iii) if $|\psi\rangle \in \text{span}\{|i\rangle|p_i\rangle\}_i + \text{span}\{|p_j^*\rangle|j\rangle\}_j$ and is orthogonal to $|\pi\rangle$, then $|||[R(P) + \mathbb{1}]|\psi\rangle|0\rangle^{\otimes ks}|| \leq 2^{1-k}$.

By the approximate reflection theorem above, there exists a subroutine $R(P)(q, k)$, where s from the statement of Theorem 3 is taken as $\log_2(q)$, and k explicitly controls the fidelity of the reflection. Note that, in the definition of the quantum reflecting agent from the main text, the parameter q is chosen in $\tilde{O}(1/\sqrt{\delta})$, and since, by a theorem of Szegedy [23], as we have commented, it holds that $1/\sqrt{\delta} \in O(1/\Delta)$, we have that the fidelity of the approximation reflection approaches unity exponentially quickly in k. We note that the parameter k should be additionally increased by a logarithmic factor of $O(\log(1/\sqrt{\epsilon}))$, in order to compensate for the accumulated error stemming from the iterations of the ARO operator, which, as clarified, we omit in this analysis. For the explicit construction of the approximate reflection operators, we refer the reader to Ref. [24].

APPENDIX C: THE PS MODEL

The PS model is a reinforcement-learning-agent model; thus, it formally fits within the specification provided with Definition 1. Here, we will recap the standard PS model introduced in Ref. [19] but note that the philosophy of the projective-simulation-based agents is not firmly confined to the formal setting we provide here, as it is more general. PS agents are defined on a more conceptual level as agents whose internal states represent episodic and compositional memory and whose deliberation comprises association-driven hops between memory sequences—so-called clips. Nonetheless, the formal definitions we give here allow us to precisely state our main claims. Following the basic definitions, we provide a formal treatment of a slight generalization of the standard

PS model which subsumes both the standard and the reflecting-agent models we refer to in the main text and which we formally treat later in this Appendix.

The PS model comprises the percept and action spaces, as given in Definition 1. The central component of PS is the so-called ECM, and it comprises the internal states of the agent. The ECM is a directed weighted network (formally represented as a directed weighted graph), the vertices of which are called clips.

Each clip c represents fragments of episodic experiences, which are formally tuples

$$c = (c^{(1)}, c^{(2)}, ..., c^{(L)}), \tag{C1}$$

where each c_k is an internal representation of a percept or an action, so

$$c^{(k)} = \mu(i), \qquad i \in \mathcal{S} \cup \mathcal{A}, \tag{C2}$$

where μ is a mapping from real percepts and actions to the internal representations. We will assume that each ECM always contains all the unit-length clips denoting elementary percepts and actions.

Within the ECM, each edge between two clips c_i and c_j is assigned a weight $h(c_i, c_j) \geq 1$ and the weights are collected in the so-called h matrix. The elementary process of the PS agent is a Markov chain, in which the excitations of the ECM hop from one clip to another, where the transition probabilities are defined by the h matrix

$$p(c_j|c_i) = \frac{h(c_i, c_j)}{\sum_k h(c_i, c_k)}; \tag{C3}$$

thus, the h matrix is just the non-normalized transition matrix. In the standard PS model, the decision function is realized as follows: Given a percept s, the corresponding clip in the ECM is excited and hopping according to how the ECM network is commenced. In the simplest case, the hopping process is terminated once a unit-length action clip is encountered, and this action is coupled out and output by the actuator (see Fig. 1). The moment when an action is coupled out can be defined in a more involved way, as we explain presently.

Finally, the update rule in the standard model necessarily involves the redefinition of the weights in the h matrix. A prototypical update rule, for a fully classical agent, defining an update from external time step t to $t + 1$ depends on whether an action has been rewarded. If the previous action has been rewarded, and the transition between clips c_i and c_j has actually occurred in the hopping process, then the update is as follows:

$$h^{(t+1)}(c_i, c_j) = h^{(t)}(c_i, c_j) - \gamma[h^{(t)}(c_i, c_j) - 1] + \lambda, \tag{C4}$$

where $0 < \lambda$ is a positive reward and $0 \leq \gamma \leq 1$ is a dissipation (forgetfulness) parameter. If the action has

not been rewarded, or the clips c_i and c_j have not played a part in the hopping process, then the weights are updated as follows:

$$h^{(t+1)}(c_i, c_j) = h^{(t)}(c_i, c_j) - \gamma[h^{(t)}(c_i, c_j) - 1]. \quad \text{(C5)}$$

The update rule can also be defined such that the update only requires the initial and terminal clips of the hopping process, which is always the case in the simple PS model, where all the clips are just actions or percepts, and hopping always involves a transition from a percept to an action. This example was used in Appendix E. For that particular example, the update function can be exactly defined by the rules above. As mentioned, aside from the basic structural and diffusion rules, the PS model allows for additional structures, which we repeat here. (i) Emoticons—the agent's short-term memory, i.e., flags that notify the agent whether the currently found action, given a percept, was previously rewarded or not. For our purposes, we shall use only the very rudimentary mode of flags that designate that the particular action (given a particular percept) was not already unsuccessfully tried before. If it was, the agent can "reflect on its decision" and reevaluate its strategies by restarting the diffusion process. This reflection process is an example of a more complicated outcoupling rule we have mentioned previously. (ii) Edge and clip glow— mechanisms that allow for the establishing of additional temporal correlations. (iii) Clip composition—the PS

model based on episodic and compositional memory allows the creation of new clips under certain variational and compositional principles. These principles allow the agent to develop new behavioral patterns under certain conditions and allow for a dynamic reconfiguration of the agent itself. For more details, we refer the reader to Refs. [19,53,54].

As illustrated, the PS model allows for a great flexibility. A straightforward generalization would allow for the ECM network to be percept specific, which is the view we adopt in the definition of reflecting agents. However, the same notion can be formalized without introducing multiple networks (one for every percept). In particular, the ECM network allows for action and percept clips to occur multiple times. Thus, the ECM network can be represented as $|\mathcal{S}|$ disjoint subnetworks, each of which comprises all elementary action clips and only one elementary percept clip. This structure is clearly within the standard PS model, and it captures all the features of the reflecting PS-agent model. A simple case of such a network, relative to the standard picture, is illustrated in Fig. 3(b). Thus, the reflecting PS-agent model is, structurally, a standard PS model as well.

APPENDIX D: BEHAVIORAL EQUIVALENCE OF CLASSICAL AND QUANTUM REFLECTING AGENTS

In this section, we show that the classical and quantum reflecting agents (RPS), denoted as A_{RPS}^c and A_{RPS}^q, are approximately equal, that is, $A_{\text{RPS}}^c \approx_\alpha A_{\text{RPS}}^q$, where α can be made arbitrarily small without incurring a significant overhead in the internal time of the agents.

We do so by separately showing that the output distributions of both the classical and the quantum RPS are α close to the previously mentioned tailed distribution, for an arbitrarily small α. The main claim will then follow by the triangle inequality on the behavioral distance (which holds since the behavioral distance is the variational distance on the output distributions).

For completeness, we begin by explicitly giving the deliberation process of the classical RPS, described in the main text, and proceed with the behavioral theorem for classical RPS.

The agent's decision-making process (implementing the decision function $\mathcal{D}$; see Appendix A for details), given percept s, is given by the following steps.

Let $t_{\text{mix}}^c \in \tilde{O}(1/\delta_s)$: (i) Sample y from some fixed distribution π_0 and (ii) repeat: (a) diffuse: (re)mix the Markov chain by

$$\pi = P_s^{t_{\text{mix}}^c} y$$

and (b) check: sample y from π. If y is a flagged action, break and output y.

In the following, when π and π' are distributions, then $\|\pi - \pi'\|$ denotes the standard variational distance (Kolmogorov distance) on distributions, so
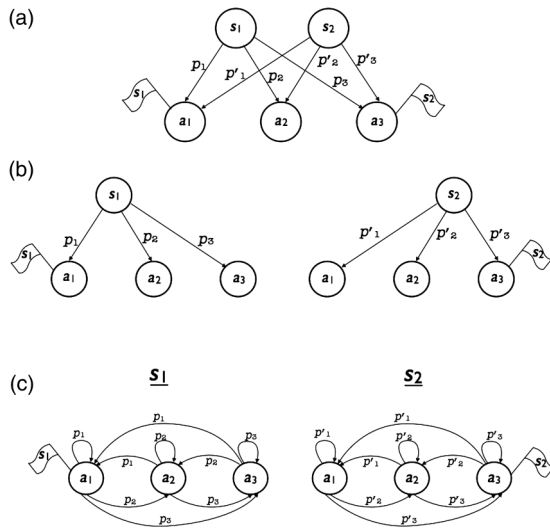

FIG. 3. (a) The simple PS model with flags. The transition from percept to action occurs in one step. Flags are percept dependent. (b) Intermediary step: The graph of (a) can be broken up into two graphs, by duplication of actions. (c) The straightforward analog of the simple PS in the reflecting-agent model. To each percept, a MC over the actions only is assigned, such that the stationary distribution recovers the initial output probabilities. The flags are trivially inherited.

$$\|\pi - \pi'\| = \frac{1}{2}\sum_x |\pi(x) - \pi'(x)|.$$

Theorem 4 (Behavior of classical reflecting agents).—
Let P_s be the transition matrix of the Markov chain associated with percept s, and let $f(s)$ be the (nonempty) set of flagged-action clips. Furthermore, let $\pi_s(x)$ be the probability mass function of the stationary distribution π_s of P_s and let $\tilde{\pi}_s$ be the renormalized distribution of π_s, where the support is retained only over the flagged actions, so

$$\tilde{\pi}_s(i) = \begin{cases} \frac{\pi_s(i)}{\sum_{j\in f(s)} \pi_s(j)} & \text{if } i \in f(s) \\ 0 & \text{otherwise.} \end{cases} \tag{D1}$$

Let κ be the probability distribution over the clips as outputted by the classical reflecting agent, upon receiving s. Then, the distance $\|\kappa - \tilde{\pi}_s\|$ is constant (ignoring logarithmic factors) and can be efficiently made arbitrarily small.

Proof.—Note that, since the Markov chain is regular, by Theorem 1, the distribution $P_s^{O(1/\delta_s)}\pi_0$, for any initial distribution π_0, is arbitrarily close to the stationary distribution of P_s; more precisely,

$$\|P_s^{[k_0+\log(1/\epsilon')/\delta_s]}\pi_0 - \pi\| \leq \epsilon' \tag{D2}$$

or, equivalently,

$$\|P_s^{(k_0+k_1)/\delta_s}\pi_0 - \pi\| \leq e^{-k_1}, \tag{D3}$$

where $k_0 = \max_i \log[\pi(i)^{-1}]$. While $\pi(i)^{-1}$ can, in principle, be very large, it only logarithmically contributes to the overhead, so it can be effectively bounded and omitted from the analysis. Thus, we can achieve an exponentially good approximation of the stationary distribution with $\tilde{O}(1/\delta_s)$ iterations. In the remainder, we will denote $\pi' = P_s^{\tilde{O}(1/\delta_s)}\pi_0$.

The reflecting agent mixes its Markov chain (achieving π') and then samples iteratively from this distribution until a flagged action is hit. Thus, $\kappa = \tilde{\pi}'$, where $\tilde{\pi}'$ is the tailed π' distribution in the sense of Eq. (D1) (substituting π_s with π').

Hence, we have that

$$\|\kappa - \tilde{\pi}_s\| = \|\tilde{\pi}' - \tilde{\pi}_s\|. \tag{D4}$$

Next, we only need to bound $\|\tilde{\pi}' - \tilde{\pi}_s\|$. Note that

$$\|\tilde{\pi}' - \tilde{\pi}_s\| = \|\frac{1}{\epsilon_{\pi'}}\pi'_{\text{sub}} - \frac{1}{\epsilon}(\pi_s)_{\text{sub}}\|, \tag{D5}$$

where ϵ and $\epsilon_{\pi'}$ are the probabilities of sampling a flagged action from π_s and π', respectively, and π'_{sub} and $(\pi_s)_{\text{sub}}$ are the subnormalized distributions $\pi'_{\text{sub}} = \epsilon_{\pi'}\tilde{\pi}'$ and $(\pi_s)_{\text{sub}} = \epsilon\tilde{\pi}_s$. Note that it holds that $\|\pi'_{\text{sub}} - (\pi_s)_{\text{sub}}\| \leq \|\pi' - \pi_s\|$. So, we have that

$$\|\frac{1}{\epsilon_{\pi'}}\pi'_{\text{sub}} - \frac{1}{\epsilon}(\pi_s)_{\text{sub}}\| = \|\frac{1}{\epsilon_{\pi'}}\pi'_{\text{sub}} - \frac{1}{\epsilon}\pi'_{\text{sub}} + \frac{1}{\epsilon}\pi'_{\text{sub}} - \frac{1}{\epsilon}(\pi_s)_{\text{sub}}\| \leq \left|\frac{1}{\epsilon_{\pi'}} - \frac{1}{\epsilon}\right| \|\pi'_{\text{sub}}\| + \frac{1}{\epsilon}\|\pi'_{\text{sub}} - (\pi_s)_{\text{sub}}\|$$

$$\leq \left|\frac{1}{\epsilon_{\pi'}} - \frac{1}{\epsilon}\right|\epsilon_{\pi'} + \frac{1}{\epsilon}\|\pi' - \pi_s\| \leq \frac{|\epsilon - \epsilon_{\pi'}|}{\epsilon} + \frac{e^{-k_1}}{\epsilon}. \tag{D6}$$

Next, note that $e^{-k_1} \geq \|\pi' - \pi_s\| \geq |\epsilon_{\pi'} - \epsilon|$, so finally, we have

$$\|\tilde{\pi}' - \tilde{\pi}_s\| \leq 2\frac{e^{-k_1}}{\epsilon} = 2e^{-k_1+\log(1/\epsilon)}. \tag{D7}$$

Since we are interested in the analysis at the $\tilde{O}$ level, we can omit the logarithmically contributing factor above. This concludes the proof. ∎

Note that since the ideal stationary distribution π (of the MC P_s) and the approximation realized by mixing are no further than e^{-k_1} apart, and since the expected time of producing an action when sampling from the ideal distribution is $O(1/\epsilon_s)$, the expected number of checks the classical RPS will perform is in $\tilde{O}(1/\epsilon_s)$.

Next, we prove an analogous theorem for the quantum agents.

Theorem 5 (Quantum reflecting agents).—Let P_s be the transition matrix of the Markov chain associated with percept s, and let $f(s)$ be the (nonempty) set of flagged-action clips. Furthermore, let $\pi_s(i)$ be the probability mass function of the stationary distribution π_s of P_s and let $\tilde{\pi}_s$ be the renormalized distribution of π_s, where the support is retained only over the flagged actions, so

$$\tilde{\pi}_s(i) = \begin{cases} \frac{\pi_s(i)}{\sum_{j\in f(s)} \pi_s(j)} & \text{if } i \in f(s) \\ 0 & \text{otherwise.} \end{cases} \tag{D8}$$

Let κ be the probability distribution over the clips as output by the quantum reflecting agent, upon receiving s. Then, the distance $\|\kappa - \tilde{\pi}\|$ is constant (up to logarithmically contributing terms) and can be efficiently made arbitrarily small.

Proof.—The proof consists of two parts. First, we prove that, if the reflection operators (used in the diffusion step of the quantum agent) are ideal, then the claim follows. The remainder of the proof, which considers imperfect reflection operators that are actually used, follows from the proof of Theorem 7 in Ref. [55]. In the following, we will, by abuse of notation, denote with $|\pi_s\rangle$ both the coherent encoding of the stationary distribution and the state $|\pi_{\text{init}}\rangle$, which is the initial state of the quantum agent and also the state over which reflections occur, following the notation of Ref. [55]. Note that the latter state is obtained by applying the quantum diffusion operator U_P once to $|\pi_s\rangle|0\rangle$, and the statistics of measuring the first register in $|\pi_{\text{init}}\rangle$ match the statistics of measuring $|\pi_s\rangle$ in the computational basis.

Assuming that the procedure the agent follows starts from the perfect stationary distribution, and that the reflections over $|\pi_s\rangle$ are then perfect, by Lemma 1, we have that the state of the system never leaves the span of $|\tilde{\pi}_s\rangle$ and $|\tilde{\pi}_s^{\perp}\rangle$, where $|\tilde{\pi}_s^{\perp}\rangle$ is the state orthogonal to $|\tilde{\pi}_s\rangle$ in span$\{|\tilde{\pi}_s\rangle, |\pi_s\rangle\}$.

Note that the agent outputs an action only conditional on the result being a flagged action. This process is equivalent to first projecting the state of the system onto the subspace of flagged actions, using the projector $\Pi^{f(s)} = \sum_{x \in f(s)} |x\rangle\langle x|$, followed by renormalization and a measurement. The state after this projection is, for any state in span$\{|\tilde{\pi}_s\rangle, |\pi_s\rangle\}$, a (sub)normalized state $|\tilde{\pi}_s\rangle$; hence, after normalization, measurement outcomes always follow the distribution $\tilde{\pi}_s$.

However, any state in span$\{|\tilde{\pi}_s\rangle, |\pi_s\rangle\}$ that is not exactly $|\tilde{\pi}_s\rangle$ still has a nonzero support on the nonflagged clips. However, it is the key feature of Grover-like search algorithms (like the search algorithm in Ref. [55], the reflections of which make up the decision process of the quantum agent) that the reflection iterations produce a state that has a constant overlap with the target state (in our case, $|\tilde{\pi}_s\rangle$, which has support only over flagged actions). This property implies that the probability of failing to hit a flagged action is, at most, some constant $\beta < 1$. Now, if the deliberation process is iterated some constant k_3 number of times, it is straightforward to see that $\|\kappa - \tilde{\pi}_s\| \leq \beta^{k_3}$ thus decays exponentially quickly, as desired.

Next, we need to consider the errors induced by the approximate reflections.

The analysis in Ref. [55] (proof of Theorem 7) directly shows that the terminal quantum (after $t_2' \in [0, t_2]$ iterations) state $|\psi\rangle$ is close to the state $|\phi\rangle$ the algorithm would produce had the reflections been perfect, which is formally

$$\| |\psi\rangle - |\phi\rangle \| \leq 2^{1-c}, \tag{D9}$$

where c is a constant only additively increasing the internal parameter k of the approximate reflection subroutine.

Thus, the error on the final state produced by the quantum agent induced by the approximate reflection algorithm can be made arbitrarily small within the allowed cost $\tilde{O}(1/\sqrt{\delta_s})$, which also means that all distributions obtained by measurements of these states will not differ by more than $4 \times 2^{1-c}$ [56].

Returning to the inequality proven for the perfect reflections, we see that the error of magnitude $4 \times 2^{1-c}$ can only increase the probability of failing to output a flagged action from β to $4 \times 2^{1-c} + \beta$. Thus, by only linearly tuning the parameters, we can make sure that the agent produces an action with an exponentially high probability in terms of the parameter k_3 (the number of iterations of the deliberation process). If an action has been produced, then, as we have shown above, it holds that it has been sampled from a distribution within a $4 \times 2^{1-c}$ distance of the tailed distribution $\tilde{\pi}_s$, which again can be efficiently made exponentially small. Thus, the theorem holds. ∎

A corollary of the two theorems above is that the classical and quantum agents are passively approximately equal. But then, from the definitions of these embodied agents, we can see that the quantum agent exhibits a consistent quadratic speedup, in terms of the required number of applications of their elementary operations. That is, the quantum agent is quadratically faster, as claimed. This proves the main claim of this paper.

The quantum agent may be required to reiterate the deliberation procedure if the final measurement finds a nonaction clip. This procedure is easy if multiple copies of the initial state are always available. Alternatively, a way of recreating the desired initial coherent encoding of the stationary distribution can be achieved by "inverting" the quantum search algorithm in Ref. [24] by "unsearching" the found nonaction clip—that is, by inverting the (unitary) Grover iterations that would be applied to it in order to search for the found clip from the initial state. This process will, with high fidelity, again recreate a good approximation of the initial stationary distribution [57] but may be costly if the found nonaction clip has a low frequency in the stationary distribution. A more efficient resolution to the problem of reiteration of deliberation can be achieved if the quantum agent can, additionally, perform a generalized measurement, mathematically described by a positive operator-valued measure (POVM), that projects the state produced by the Grover-like iterations to the subspaces spanned by nonflagged and flagged clips, defined by the POVM elements

$$\Pi^{f(s)} = \sum_{i \in f(s)} |i\rangle\langle i|, \tag{D10}$$

$$\Pi^{\text{non-}f(s)} = \mathbb{1} - \Pi^{f(s)}. \tag{D11}$$

If the first outcome of this measurement is obtained, the agent will output the required action according to the desired distribution, by measuring the residual state. If

the other outcome is obtained, then the resulting state is the coherent encoding of the stationary distribution, renormalized so that it has support only over nonflagged clips. Starting from this state (which is in the work space of the Grover-like iterations), the agent can "unsearch" for all nonflagged clips to regain the initial distribution. Crucially, the number of steps here will scale as $\tilde{O}(1/\sqrt{1-\epsilon_s})$, which will, in interesting cases, be effectively constant.

APPENDIX E: COMPARISON OF PS-AGENT MODELS

In this section, we compare reflecting PS agents with the standard PS agents, which have been previously studied [19,53]. The main difference between the standard and the reflecting models is that the standard agents evolve their Markov chain until the first instance where an action clip has been hit, whereas the reflecting agents allow for the Markov chain to fully mix. In the standard PS model, the underlying Markov chain should never mix, as indeed, in that case, the construction will guarantee that the actual action performed is independent from the received percept (if the Markov chain is irreducible). This observation suggests that large clip networks of high conductivity that could, e.g., appear in agents that have undergone a rich complex set of episodic experiences (for instance, any type of clip compositions we mentioned previously) are better suited for the reflecting, rather than the standard, PS agents.

Conversely, since the reflecting agent must always fully mix its Markov chain, and the standard does not, it would seem that the standard model should typically outperform the reflecting model on simple clip networks (as the hitting time may be much smaller than the mixing time). This intuition would suggest that the standard and reflecting PS models should not be compared on simple structures, as no quantum advantage can be demonstrated.

In contrast to these intuitive hypotheses, here, we show that even for the simplest (but nontrivial) standard PS-agent construction [19,53], there exists a natural reflecting-agent analog, where the performance of the classical reflecting agent matches the standard PS construction, and consequently, the quantum reflecting agent yields a quadratic speedup over both.

1. Simple standard PS agent with flags

Recall that, in the standard PS model, the internal state of the agent is defined by an h matrix, which defines transition probabilities of the directed Markov chain over the clip space. In the simplest case, the clip space comprises only percepts and actions, the agent is initialized so that each percept is connected to each action with unit weight (implying equiprobable transitions), and no other connections exist. Furthermore, the update function we assume is the standard update function, as defined in Ref. [19] and repeated in Appendix C. In this particular model, the internal elementary time step is one evolution of the

Markov chain, so the agent always decides in one step, and there is nothing to improve. However, if we introduce short-term memory, which significantly improves the performance of the model [19], then the fact that the agent may hit an unflagged item implies it may have to walk again. If ϵ_s denotes the probability (for the clip s) that the agent hits a flagged item, then the expected number of elementary transitions and checks the agent will perform is $O(1/\epsilon_s)$. (See Fig. 3 for an illustration.) For this model, we next introduce a direct reflecting-agent analog.

2. Simple reflecting agent with flags

It is relatively straightforward to construct a reflecting PS agent that is behaviorally equivalent to the simple standard agent with flags, given above. To each percept, we assign a Markov chain over just the action space, such that the transition matrix is column constant and the column simply contains the transition probabilities for that particular percept in the standard model above; see Fig. 3 for an illustration. The stationary distribution of this Markov chain then matches the desired transition probabilities. The update function for this agent is then exactly the same as in the standard model (that is, internally, the reflecting agent can also keep and update the same h matrix that induces the Markov chains), and the flags are also treated in exactly the same way.

Note that the transition matrix of such a Markov chain is rank one, meaning it has only one nonzero eigenvalue, and since the trace of this matrix is unit (and trace is invariant under diagonalization or basis change), the largest eigenvalue must be 1 and all the others are 0. Thus, we have that the spectral gap is always 1. This fact immediately implies that the Markov chain is fully mixed after just a single transition (as $\delta = 1$). Thus, the classical reflecting agent will also perform $O(1/\epsilon_s)$ transitions and checks until an action is performed, given the percept s. However, the quantum agent only requires $O(1/\sqrt{\epsilon_s})$ calls to the quantum-walk operator. Thus, even in the very simple setting, we see that the reflecting agents can be compared to the standard PS-agent model and maintain a generic quadratic speedup when quantized.

This simple model also serves as an illustration of a setting in which the initial state $|\pi_s\rangle$ is easily prepared. Since the relevant transition matrix P_s is of rank one, it is straightforward to see that one application of the quantum diffusion operator U_{P_s} applied on the register initialized to $|1\rangle|0\rangle$ generates the state $|1\rangle \otimes \sum_j \sqrt{P_{j,1}}|j\rangle = |1\rangle|\pi_s\rangle$, since the P_s is a column-constant matrix containing the stationary distribution as columns.

While the scenario considered in this section is somewhat restricted, it is not without importance, as, for the standard PS agents, we already have a body of results in the classical case [19,53]. We emphasize, however, that the quadratic speedup proven in this work is not restricted to this scenario.

[1] D. Deutsch, *Quantum Theory, The Church-Turing Principle and the Universal Quantum Computer*, Proc. R. Soc. A **400,** 97 (1985).

[2] D. Deutsch and R. Jozsa, *Rapid Solution of Problems by Quantum Computation*, Proc. R. Soc. A **439,** 553 (1992).

[3] L. K. Grover, *Proceedings of the 28th Annual ACM Symposium on the Theory of Computing (STOC)* (ACM, New York, 1996), pp. 212–219.

[4] P. W. Shor, in *Proceedings of the 35th Symposium on Foundations of Computer Science, FOCS 1994* (IEEE, New York, 1994), p. 124–134.

[5] M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information* (Cambridge University Press, Cambridge, England, 2000).

[6] C. H. Bennett and D. P. DiVincenzo, *Quantum Information and Computation*, Nature (London) **404,** 247 (2000).

[7] M. Sasaki and A. Carlini, *Quantum Learning and Universal Quantum Matching Machine*, Phys. Rev. A **66,** 022303 (2002).

[8] H. Neven, V. S. Denchev, G. Rose, and W. G. Macready, *Training a Binary Classifier with the Quantum Adiabatic Algorithm*, arXiv:0811.0416.

[9] S. Lloyd, M. Mohseni, and P. Rebentrost, *Quantum Algorithms for Supervised and Unsupervised Machine Learning*, arXiv:1307.0411.

[10] D. Manzano, M. Pawłowski, and Č. Brukner, *The Speed of Quantum and Classical Learning for Performing the kth Root of Not*, New J. Phys. **11,** 113018 (2009).

[11] K. L. Pudenz and D. A. Lidar, *Quantum Adiabatic Machine Learning*, Quantum Inf. Process. **12,** 2027 (2013).

[12] E. Aïmeur, G. Brassard, and S. Gambs, *Quantum Speed-up for Unsupervised Learning*, Mach. Learn. **90,** 261 (2013).

[13] V. Braitenberg, *Vehicles: Experiments in Synthetic Psychology* (MIT Press, Cambridge, MA, 1986).

[14] R. Brooks, *Cambrian Intelligence: The Early History of the New AI* (MIT Press, Cambridge, MA, 1999).

[15] R. Pfeifer and C. Scheier, *Understanding Intelligence* (MIT Press, Cambridge, MA, 1999).

[16] R. Pfeifer and J. C. Bongard, *How the Body Shapes the Way We Think: A New View of Intelligence* (MIT Press, Cambridge, MA, 2006).

[17] D. Floreano and C. Mattiussi, *Bio-inspired Artificial Intelligence: Theories, Methods, and Technologies* (MIT Press, Cambridge, MA, 2008).

[18] L. W. Barsalou, *Grounded Cognition*, Annu. Rev. Psychol. **59,** 617 (2008).

[19] H. J. Briegel and G. De las Cuevas, *Projective Simulation for Artificial Intelligence*, Sci. Rep. **2,** 400 (2012).

[20] S. J. Russel and P. Norvig, *Artificial Intelligence—A Modern Approach* (Prentice-Hall, New Jersey, 2003), 2nd ed.

[21] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction* (MIT Press, Cambridge, MA, 1998), 1st ed.

[22] AIBO, adapted from http://commons.wikimedia.org/wiki/Category:Aibo#mediaviewer/File:Aibo_img_0807.jpg, created by Rama, and licensed under the CC-BY-SA-2.0-FR.

[23] M. Szegedy, in *Proceedings of the 45th Annual IEEE Symposium on Foundations of Computer Science, 2004* (IEEE Computer Society, Washington, DC, 2004), pp. 32–41.

[24] F. Magniez, A. Nayak, J. Roland, and M. Santha, *Search via Quantum Walk*, SIAM J. Comput. **40,** 142 (2011).

[25] Examples of such applications include the problems of navigation, or human-robot interaction. We note that the focus on such classical environments is by no means a restriction of our scheme. For a quantum environment, the actions that the agent can perform could, e.g., be quantum measurements (as components of his actuators), and percepts the measurement outcomes—such scenarios are certainly not without interest. The results we present in this work apply equally to such environments. However, we will not explore the applications of learning agents in quantum environments in this paper.

[26] Even if we were to allow superpositions of actions, the amount of control the agent must have over the degrees of freedom of the environment, in order to apply quantum query algorithms, may be prohibitive. This problem constitutes one of the fundamental operative distinctions between quantum algorithms, where full control is assumed, and quantum agents, where such control is limited.

[27] In embodied agents, these physical processes may, e.g., realize some internal representation of the environment, which the agent itself has to develop as it interacts with the environment. For example, in the context of artificial neural networks, such internal models are known as self-organizing maps and, more specifically, sensorimotor maps [58,59].

[28] In the PS framework, one formally distinguishes between real percepts and actions s and a and their internal representations, denoted as $\mu(s)$ and $\mu(a)$, which comprise clips. For our purposes, by abuse of notation, we will omit the mapping μ. For more details, see Appendix C.

[29] In the PS, the hopping probabilities themselves are encoded in the so-called h matrix, which is an unnormalized representation of the transition matrix [19].

[30] Each agent is equipped with input and output couplers, which translate, through sensors and actuators (see Fig. 1), real percepts to the internal representations of percepts, and internal representations of actions to real actions.

[31] In this paper, we do not consider logarithmically contributing terms in the complexity analysis; thus, we use the $\tilde{O}$-level analysis of the limiting behavior (instead of the standard "big-O" O).

[32] F. Magniez, M. Santha, and M. Szegedy, in *Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2005* (Society for Industrial and Applied Mathematics, Philadelphia, 2005), p. 1109–1117.

[33] A. Ambainis, in *Proceedings of the 45th Annual IEEE Symposium on Foundations of Computer Science, October 17–19, 2004* (IEEE Computer Society Press, Los Alamitos, CA, 2004), p. 22–31.

[34] F. Magniez, A. Nayak, P. C. Richter, and M. Santha, *On the Hitting Times of Quantum versus Random Walks*, Algorithmica **63,** 91 (2012).

[35] The optimality claim holds provided that no additional mechanisms, except for diffusion and checking, are available to the agent.

[36] K. Temme, T. J. Osborne, K. G. H. Vollbrecht, D. Poulin, and F. Verstraete, *Quantum Metropolis Sampling*, Nature (London) **471,** 87 (2011).

[37] M.-H. Yung and A. Aspuru-Guzik, *A Quantum–Quantum Metropolis Algorithm*, Proc. Natl. Acad. Sci. U.S.A. **109,** 754 (2012).

[38] R. D. Somma, S. Boixo, H. Barnum, and E. Knill, *Quantum Simulations of Classical Annealing Processes*, Phys. Rev. Lett. **101,** 130504 (2008).

[39] P. Wocjan and A. Abeyesinghe, *Speedup via Quantum Sampling*, Phys. Rev. A **78,** 042336 (2008).

[40] In the case of reversible MCs, which will be the main focus of this paper, $P = P^*$, so V can be constructed from U by conjugating it with the swap operator of the two registers. Here, we present the construction for the general case of irreducible, aperiodic Markov chains.

[41] Here, we note that concrete applications of quantum walks specified by nonsymmetric Markov chains have, to our knowledge and aside from this work, only been studied in Refs. [60,61], by two of the authors and other collaborators, in a significantly different context.

[42] For completeness, we note, as a technicality, following Ref. [24], that if ϵ_s is always bounded below by a known constant (say, 3/4, as in Ref. [43]), the quantum agent can immediately measure the initial state and efficiently produce the desired output by iterating this process a constant number of times. However, in the scenarios we envision, ϵ_s is very small.

[43] M. Boyer, G. Brassard, P. Høyer, and A. Tapp, *Tight Bounds on Quantum Searching*, Fortschr. Phys. **46,** 493 (1998).

[44] We note that some of the only logarithmically contributing terms, which appear in a more detailed analysis omitted here, can be further avoided using more complicated constructions, as has, for instance, been done in Ref. [24].

[45] A. Aspuru-Guzik and P. Walther, *Photonic Quantum Simulators*, Nat. Phys. **8,** 285 (2012).

[46] R. Blatt and C. F. Roos, *Quantum Simulations with Trapped Ions*, Nat. Phys. **8,** 277 (2012).

[47] T. Schaetz, C. R. Monroe, and T. Esslinger, *Focus on Quantum Simulation*, New J. Phys. **15,** 085009 (2013).

[48] D. H. Wolpert, *The Lack of A Priori Distinctions between Learning Algorithms*, Neural Comput. **8,** 1341 (1996).

[49] S. Droste, T. Jansen, and I. Wegener, *Optimization with Randomized Search Heuristics—The (a)nfl Theorem, Realistic Scenarios, and Difficult Functions*, Theor. Comput. Sci. **287,** 2002 (1997).

[50] We acknowledge that the interpretation of these results in the sense of their practical impact on the field is not without controversy. Nonetheless, the validity of the mathematical statements is not contested. See Ref. [62] for more details.

[51] That is, a comparison can be made with no further assumptions on the environment beyond the trivial ones —that the percept and action sets are compatible with the environment and that the environment provides a rewarding scheme.

[52] A. Sinclair, *Algorithms for Random Generation and Counting: A Markov Chain Approach* (Springer, New York, 1993), Vol. 7.

[53] J. Mautner, A. Makmal, D. Manzano, M. Tiersch, and H. J. Briegel, arXiv:1305.1578. To appear in New Generation Computing (2014).

[54] H. J. Briegel, *On Creative Machines and the Physical Origins of Freedom*, Sci. Rep. **2,** 522 (2012).

[55] F. Magniez, M. Santha, and M. Szegedy, *Quantum Algorithms for the Triangle Problem*, SIAM J. Comput. **37,** 413 (2007).

[56] E. Bernstein and U. Vazirani, in *Proceedings of the 25th Annual ACM Symposium on Theory of Computing* (ACM, New York, 1993), pp. 11–20.

[57] Additionally, the quality of the approximation can be, probabilistically, arbitrarily raised by one round of the phase-estimation algorithm for the walk operator, provided that no phase is detected. Should this process fail, the unsearching process is repeated.

[58] T. Kohonen, *Self-Organizing Maps* (Springer-Verlag, Berlin, 1995).

[59] M. Toussaint, *A Sensorimotor Map: Modulating Lateral Interactions for Anticipation and Planning*, Neural Comput. **18,** 1132 (2006).

[60] G. D. Paparo and M. Martin-Delgado, *Google in a Quantum Network*, Sci. Rep. **2,** 444 (2012).

[61] G. Paparo, M. Mueller, F. Comellas, and M. Martin-Delgado, *Quantum Google in a Complex Network*, Sci. Rep. **3,** 2773 (2013).

[62] An introduction to discussions regarding the no free lunch theorems, and their interpretations, can be found at http://no-free-lunch.org/.