SCIENTIFIC REPORTS

OPEN

α-Rank: Multi-Agent Evaluation by Evolution

Shayegan Omidshafiei[1], Christos Papadimitriou[5], Georgios Piliouras[4], Karl Tuyls[1], Mark Rowland[2], Jean-Baptiste Lespiau[1], Wojciech M. Czarnecki[2], Marc Lanctot[3], Julien Perolat[2] & Remi Munos[1]

Received: 18 February 2019
Accepted: 11 June 2019
Published online: 09 July 2019

We introduce α-*Rank*, a principled evolutionary dynamics methodology, for the *evaluation* and *ranking* of agents in large-scale multi-agent interactions, grounded in a novel dynamical game-theoretic solution concept called *Markov-Conley chains* (MCCs). The approach leverages continuous-time and discrete-time evolutionary dynamical systems applied to empirical games, and scales tractably in the number of agents, in the type of interactions (beyond dyadic), and the type of empirical games (symmetric and asymmetric). Current models are fundamentally limited in one or more of these dimensions, and are not guaranteed to converge to the desired game-theoretic solution concept (typically the Nash equilibrium). α-Rank automatically provides a ranking over the set of agents under evaluation and provides insights into their strengths, weaknesses, and long-term dynamics in terms of basins of attraction and sink components. This is a direct consequence of the correspondence we establish to the dynamical MCC solution concept when the underlying evolutionary model's ranking-intensity parameter, α, is chosen to be large, which exactly forms the basis of α-Rank. In contrast to the Nash equilibrium, which is a static solution concept based solely on fixed points, MCCs are a dynamical solution concept based on the Markov chain formalism, Conley's Fundamental Theorem of Dynamical Systems, and the core ingredients of dynamical systems: fixed points, recurrent sets, periodic orbits, and limit cycles. Our α-Rank method runs in polynomial time with respect to the total number of pure strategy profiles, whereas computing a Nash equilibrium for a general-sum game is known to be intractable. We introduce mathematical proofs that not only provide an overarching and unifying perspective of existing continuous- and discrete-time evolutionary evaluation models, but also reveal the formal underpinnings of the α-Rank methodology. We illustrate the method in canonical games and empirically validate it in several domains, including AlphaGo, AlphaZero, MuJoCo Soccer, and Poker.

This paper introduces a principled, practical, and descriptive methodology, which we call α-*Rank*. α-Rank enables evaluation and ranking of agents in large-scale multi-agent settings, and is grounded in a new game-theoretic solution concept, called Markov-Conley chains (MCCs), which captures the dynamics of multi-agent interactions. While much progress has been made in learning for games such as Go[1,2] and Chess[3], computational gains are now enabling algorithmic innovations in domains of significantly higher complexity, such as Poker[4] and MuJoCo soccer[5] where ranking of agents is much more intricate than in classical simple matrix games. With multi-agent learning domains of interest becoming increasingly more complex, we need methods for evaluation and ranking that are both comprehensive and theoretically well-grounded.

Evaluation of agents in a multi-agent context is a hard problem due to several complexity factors: strategy and action spaces of players quickly explode (e.g., multi-robot systems), models need to be able to deal with intransitive behaviors (e.g., cyclical best-responses in Rock-Paper-Scissors, but at a much higher scale), the number of agents can be large in the most interesting applications (e.g., Poker), types of interactions between agents may be complex (e.g., MuJoCo soccer), and payoffs for agents may be asymmetric (e.g., a board-game such as Scotland Yard).

This evaluation problem has been studied in Empirical Game Theory using the concept of empirical games or meta-games, and the convergence of their dynamics to Nash equilibria[6–9]. In Empirical Game Theory a meta-game is an abstraction of the underlying game, which considers meta-strategies rather than primitive actions[6,8]. In the Go domain, for example, meta-strategies may correspond to different AlphaGo agents (e.g.,

[1]DeepMind, Paris, France. [2]DeepMind, London, UK. [3]DeepMind, Edmonton, Canada. [4]Singapore University of Technology and Design, Singapore, Singapore. [5]Columbia University, New York, USA. Shayegan Omidshafiei, Christos Papadimitriou, Georgios Piliouras and Karl Tuyls contributed equally. Correspondence and requests for materials should be addressed to K.T. (email: karltuyls@google.com)

each meta-strategy is an agent using a set of specific training hyperparameters, policy representations, and so on). The players of the meta-game now have a choice between these different agents (henceforth synonymous with meta-strategies), and payoffs in the meta-game are calculated corresponding to the win/loss ratio of these agents against each other over many rounds of the full game of Go. Meta-games, therefore, enable us to investigate the strengths and weaknesses of these agents using game-theoretic evaluation techniques.

Existing meta-game analysis techniques, however, are still limited in a number of ways: either a low number of players or a low number of agents (i.e., meta-strategies) may be analyzed[6–8,10]. Specifically, on the one hand continuous-time meta-game evaluation models, using replicator dynamics from Evolutionary Game Theory[11–15], are deployed to capture the micro-dynamics of interacting agents. These approaches study and visualize basins of attraction and equilibria of interacting agents, but are limited as they can only be feasibly applied to games involving few agents, exploding in complexity in the case of large and asymmetric games. On the other hand, existing discrete-time meta-game evaluation models (e.g.[16–20]) capture the macro-dynamics of interacting agents, but involve a large number of evolutionary parameters and are not yet grounded in a game-theoretic solution concept.

To further compound these issues, using the Nash equilibrium as a solution concept for meta-game evaluation in these dynamical models is in many ways problematic: first, computing a Nash equilibrium is computationally difficult[21,22]; second, there are intractable equilibrium selection issues even if Nash equilibria can be computed[23–25]; finally, there is an inherent incompatibility in the sense that it is not guaranteed that dynamical systems will converge to a Nash equilibrium[26,27], or, in fact, to any fixed point. However, instead of taking this as a disappointing flaw of dynamical systems models, we see it as an opportunity to look for a novel solution concept that does not have the same limitations as Nash in relation to these dynamical systems. Specifically, exactly as J. Nash used one of the most advanced topological results of his time, i.e., Kakutani's fixed point theorem[28], as the basis for the Nash solution concept, in the present work, we employ Conley's Fundamental Theorem of Dynamical Systems[29] and propose the solution concept of Markov-Conley chains (MCCs). Intuitively, Nash is a static solution concept solely based on fixed points. MCCs, by contrast, are a dynamic solution concept based not only on fixed points, but also on recurrent sets, periodic orbits, and limit cycles, which are fundamental ingredients of dynamical systems. The key advantages are that MCCs comprehensively capture the long-term behaviors of our (inherently dynamical) evolutionary systems, and our associated α-Rank method runs in polynomial time with respect to the total number of pure strategy profiles (whereas computing a Nash equilibrium for a general-sum game is PPAD-complete[21]).

Main Contributions: α-Rank and MCCs

The contribution of this paper is three-fold: 1) the introduction of a multi-population discrete-time model, which enables evolutionary analysis of many-player interactions even in asymmetric games, 2) the introduction of the MCC solution concept, a new game-theoretic concept that captures the dynamics of multi-agent interactions, and subsequent connection to the discrete-time model, and 3) the specific ranking strategy/algorithm for the general multi-population setting that we call α-Rank. While MCCs do not immediately address the equilibrium selection problem, we show that by introducing a perturbed variant that corresponds to a generalized multi-population discrete-time dynamical model, the underlying Markov chain containing them becomes irreducible and yields a unique stationary distribution. The ordering of the strategies of agents in this distribution gives rise to our α-Rank methodology. α-Rank provides a summary of the asymptotic evolutionary rankings of agents in the sense of the time spent by interacting populations playing them, yielding insights into their evolutionary strengths. It both automatically produces a ranking over agents favored by the evolutionary dynamics and filters out transient agents (i.e., agents that go extinct in the long-term evolutionary interactions).

Paper Overview

Due to the interconnected nature of the concepts discussed herein, we provide in Fig. 1 an overview of the paper that highlights the relationships between them. Due to the technical background necessary for fully understanding the paper contribution, we give readers the choice of a 'short' vs. 'long' read-through of the paper, with the short read-through consisting of the sections highlighted in green in Fig. 1 and suited for the reader who wants to quickly grasp the high-level ideas, and the long read-through consisting of all technical details.

Specifically, the paper is structured as follows: we first provide a review of preliminary game-theoretic concepts, including the Nash equilibrium (box ① in Fig. 1), which is a long-standing yet static solution concept. We then overview the replicator dynamics micro-model (②), which provides low-level insights into agent interactions but is limited in the sense that it can only feasibly be used for evaluating three to four agents. We then introduce a generalized evolutionary macro-model (③) that extends previous single-population discrete-time models (④) and (as later shown) plays an integral role in our α-Rank method. We then narrow our focus on a particular evolutionary macro-model (③) that generalizes single-population discrete-time models (④) and (as later shown) plays an integral role in our α-Rank method. Next, we highlight a fundamental incompatibility of the dynamical systems and the Nash solution concept (⑤), establishing fundamental reasons that prevent dynamics from converging to Nash. This limitation motivates us to investigate a novel solution concept, using Conley's Fundamental Theorem of Dynamical Systems as a foundation (⑥).

Conley's Theorem leads us to the topological concept of *chain components*, which do capture the irreducible long-term behaviors of a *continuous* dynamical system, but are unfortunately difficult to analyze due to the lack of an exact characterization of their geometry and the behavior of the dynamics inside them. We, therefore, introduce a discrete approximation of these limiting dynamics that is more feasible to analyze: our so-called Markov-Conley chains solution concept (⑦). While we show that Markov-Conley chains share a close theoretical relationship with both discrete-time and continuous-time dynamical models (⑧), they unfortunately suffer from an equilibrium selection problem and thus cannot directly be used for computing multi-agent rankings. To address this, we introduced a perturbed version of Markov-Conley chains that resolves the equilibrium selection issues and yields our α-Rank evaluation method (⑨). α-Rank computes both a ranking and assigns scores to

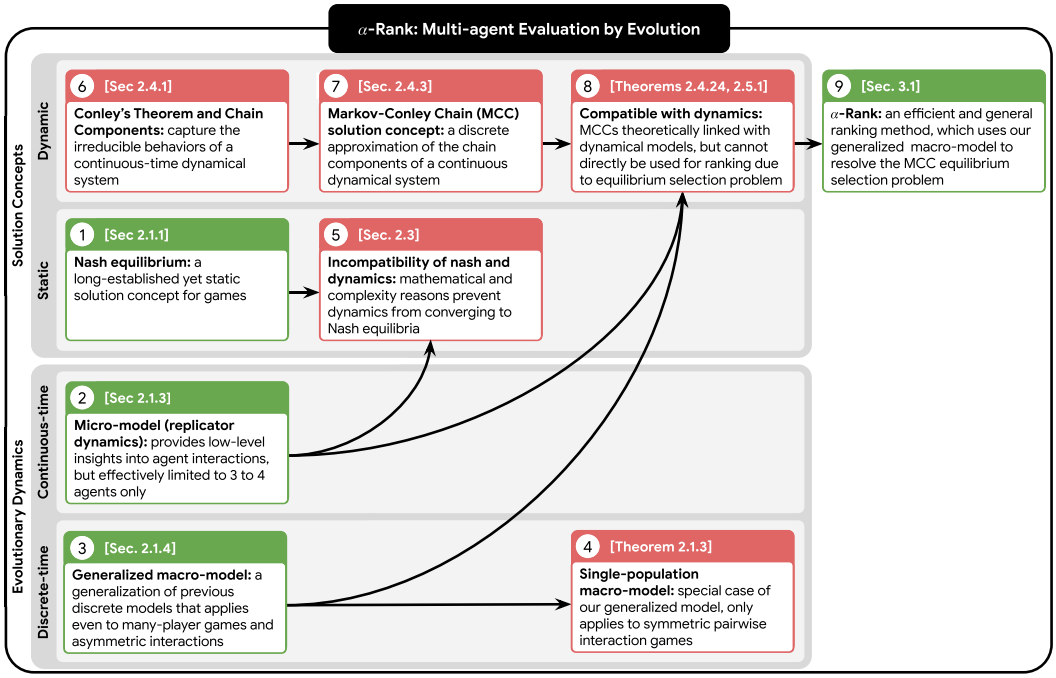


Figure 1. Paper at a glance. Numerical ordering of the concept boxes corresponds to the paper flow, with sections and/or theorems indicated where applicable. Due to the technical background necessary for fully understanding the paper contribution, we give readers the choice of a 'short' vs. 'long' read-through of the paper, with the short read-through consisting of the sections highlighted in green in this figure and suited for the reader who wants to quickly grasp the high-level ideas, and the long read-through consisting of all technical details. The methods and ideas used herein may be classified broadly as either game-theoretic *solution concepts* (namely, static or dynamic) and *evolutionary dynamics* concepts (namely, continuous- or discrete-time). The insights gained by analyzing existing concepts and developing new theoretical results carves a pathway to the novel combination of our general multi-agent evaluation method, α-Rank, and our game-theoretic solution concept, Markov-Conley Chains.

agents using this perturbed model. We show that this perturbed model corresponds directly to the generalized macro-model under a particular setting of the latter's so-called *ranking-intensity parameter* α. α-Rank not only captures the dynamic behaviors of interacting agents, but is also more tractable to compute than Nash for general games. We validate our methodology empirically by providing ranking analysis on datasets involving interactions of state-of-the-art agents including AlphaGo[1], AlphaZero[3], MuJoCo Soccer[5], and Poker[30], and also provide scalability properties and theoretical guarantees for the overall ranking methodology.

Preliminaries and Methods

In this section, we concisely outline the game-theoretic concepts and methods necessary to understand the remainder of the paper. We also introduce a novel game-theoretic concept, Markov-Conley chains, which we use to theoretically ground our results in. Readers familiar with game theory or dynamical systems may wish to, respectively, skip Sections 2.1.1 to 2.1.3 and Sections 2.4.1 and 2.4.2. For a detailed discussion of the concepts we refer the reader to[6,13,31,32].

Game theoretic concepts. *Normal form games.* A K-wise interaction Normal Form Game (NFG) G is defined as $(K, \prod_{k=1}^{K} S^k, \prod_{k=1}^{K} M^k)$, where each player $k \in \{1, \ldots, K\}$ chooses a strategy s^k from its strategy set S^k and receives a payoff M^k: $\prod_{i=1}^{K} S^i \to \mathbb{R}$. We henceforth denote the joint strategy space and payoffs, respectively, as $\prod_k S^k$ and $\prod_k M^k$. We denote the strategy profile of all players by $s = (s^1, \ldots, s^K) \in \prod_k S^k$, the strategy profile of all players except K by s^{-k}, and the payoff profile by $(M^1(s^1, s^{-1}), \ldots, M^K(s^K, s^{-K}))$. An NFG is symmetric if the following two conditions hold: first, all players have the same strategy sets (i.e., $\forall\, k, l\, S^k = S^l$); second, if a permutation is applied to the strategy profile, the payoff profile is permuted accordingly. The game is asymmetric if one or both of these conditions do not hold. Note that in a 2-player ($K = 2$) NFG the payoffs for both players (M above) are typically represented by a bi-matrix (A, B), which gives the payoff for the row player in A, and the payoff for the column player in B. If $S^1 = S^2$ and $A = B^T$, then this 2-player game is symmetric.

In the case of randomized (mixed) strategies, we typically overload notation as follows: if x^k is a mixed strategy for each player k and x^{-k} the mixed profile excluding that player, then we denote by $M^k(x^k, x^{-k})$ the expected payoff of player k, $\mathbf{E}_{s^k \sim x^k, s^{-k} \sim x^{-k}}[M^k(s^k, s^{-k})]$. Given these preliminaries, we are now ready to define the Nash equilibrium concept:

Concept	Meaning
K-wise meta-game	An NFG with K player slots.
Strategy	The agents under evaluation (e.g., variants of AlphaGo agents) in the meta-game.
Individual	A population member, playing a strategy and assigned to a slot in the meta-game.
Population	A finite set of individuals.
Player	An individual that participates in the meta-game under consideration.
Monomorphic Population	A finite set of individuals, playing the same strategy.
Monomorphic Population Profile	A set of monomorphic populations.
Focal Population	A previously-monomorphic population wherein a rare mutation has appeared.

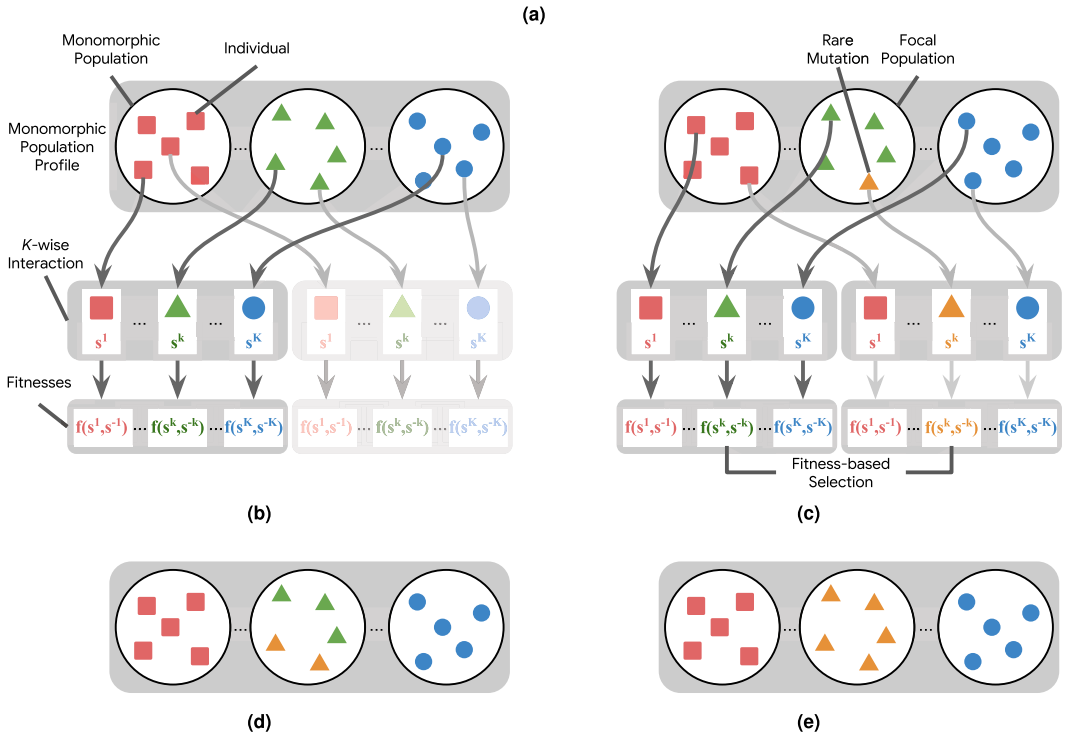


Figure 2. Overview of the discrete-time macro-model. (**a**) Evolutionary concepts terminology. (**b**) We have a set of individuals in each population k, each of which is programmed to play a strategy from set S^k. Under the mutation rate $\mu \to 0$ assumption, at most one population is not monomorphic at any time. Each individual in a K-wise interaction game has a corresponding fitness $f^k(s^k, s^{-k})$ dependent on its identity k, its strategy s^k, and the strategy profile s^{-k} of the other players. (**c**) Let the focal population denote a population k wherein a rare mutant strategy appears. At each timestep, we randomly sample two individuals in population k; the strategy of the first individual is updated by either probabilistically copying the strategy of the second individual, mutating with a very small probability to a random strategy, or sticking with its own strategy. (**d**) Individual in the focal population copies the mutant strategy. (**e**) The mutant propagates in the focal population, yielding a new monomorphic population profile.

Definition 2.1.1 (Nash equilibrium). A mixed strategy profile $x = (x^1, \ldots, x^K)$ is a Nash equilibrium if for all players k: $\max_{x'^k} M^k(x'^k, x^{-k}) = M^k(x^k, x^{-k})$.

Intuitively, a strategy profile x is a Nash equilibrium of the NFG if no player has an incentive to unilaterally deviate from its current strategy.

Meta-games. A meta-game (or an empirical game) is an NFG that provides a simplified model of an underlying multi-agent system (e.g., an auction, a real-time strategy game, or a robot football match), which considers meta-strategies or 'styles of play' of agents, rather than the full set of primitive strategies available in the underlying game[6,8,9]. Empirical (or meta-) games will play an instrumental role in our endeavor. Note also that a different notion of meta-games is discussed in earlier work[33], but plays no role here. In this paper, the meta-strategies considered are learning agents (e.g., different variants of AlphaGo agents, as exemplified in Section 1). Thus, we henceforth refer to meta-games and meta-strategies, respectively, as 'games' and 'agents' when the context is clear. For example, in AlphaGo, styles of play may be characterized by a set of agents $\{AG(r), AG(v), AG(p)\}$, where AG stands for the algorithm and indexes r, v, and p stand for *rollouts*, *value networks*, and *policy networks*, respectively, that lead to different play styles. The corresponding meta-payoffs quantify the outcomes when players play profiles over the set of agents (e.g., the empirical win rates of the agents when played against one another). These payoffs can be calculated from available data of the agents' interactions in the real multi-agent systems (e.g., wins/losses in the game of Go), or they can be computed from simulations. The question of how many such interactions

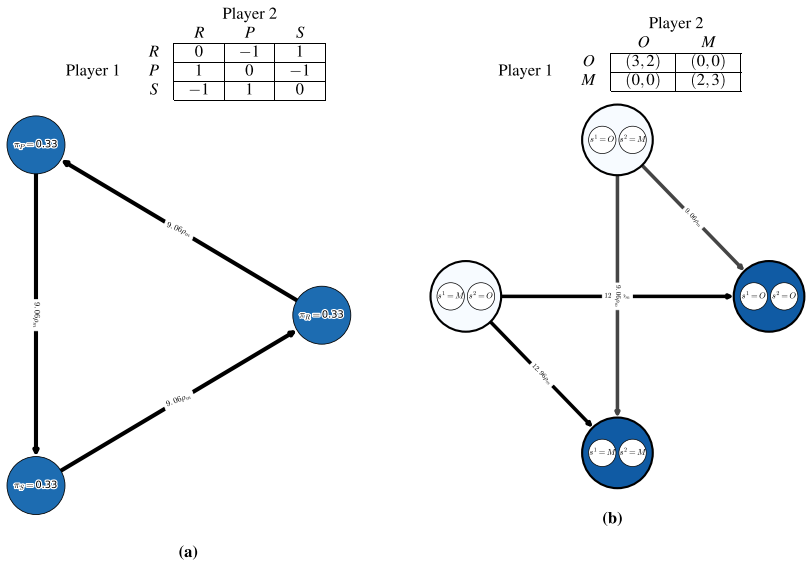


Figure 3. Conceptual examples of finite-population models, for population size $m = 50$ and ranking-intensity $\alpha = 0.1$. **(a)** Payoffs (top) and single-population discrete-time dynamics (bottom) for Rock-Paper-Scissors game. Graph nodes correspond to monomorphic populations R, P, and S. **(b)** Payoffs (top) and multi-population discrete-time dynamics (bottom) for Battle of the Sexes game. Strategies O and M respectively correspond to going to the Opera and Movies. Graph nodes correspond to monomorphic population profiles (s1; s2). The stationary distribution p has 0.5 mass on each of profiles (O;O) and (M;M), and 0 mass elsewhere.

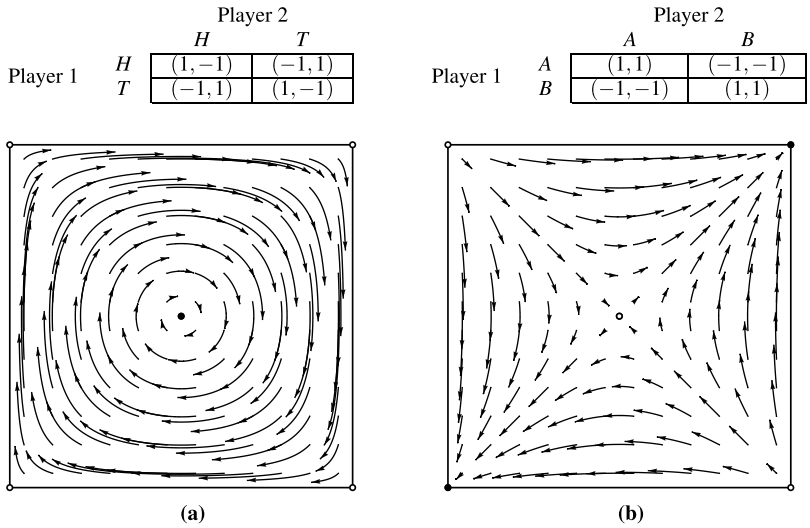


Figure 4. Canonical game payoffs and replicator dynamics trajectories. Each point encodes the probability assigned by the players to their first strategy. The matching pennies replicator dynamics have *one* chain component, consisting of the whole domain. The coordination game dynamics have five chain components (corresponding to the fixed points, four in the corners and one mixed, which are recurrent by definition), as was formally shown by[26].

are necessary to have a good approximation of the true underlying meta-game is discussed in[6]. A meta-game itself is an NFG and can, thus, leverage the game-theoretic toolkit to evaluate agent interactions at a high level of abstraction.

Micro-model: replicator dynamics. *Dynamical systems* is a powerful mathematical framework for specifying the time dependence of the players' behavior (see the Supplementary Material for a brief introduction).

For instance, in a two-player asymmetric meta-game represented as an NFG $(2, S^1 \times S^2, M = (A, B))$, the evolution of players' strategy profiles under the replicator dynamics[34,35] is given by,

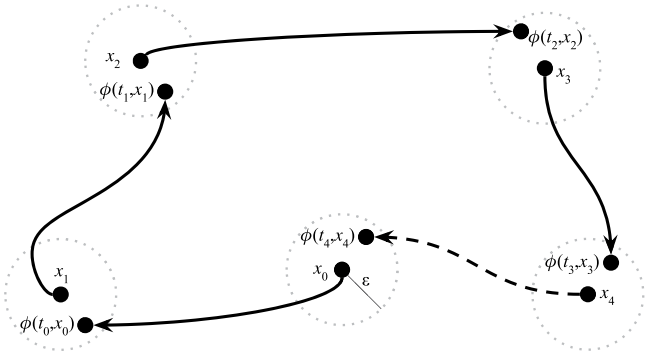
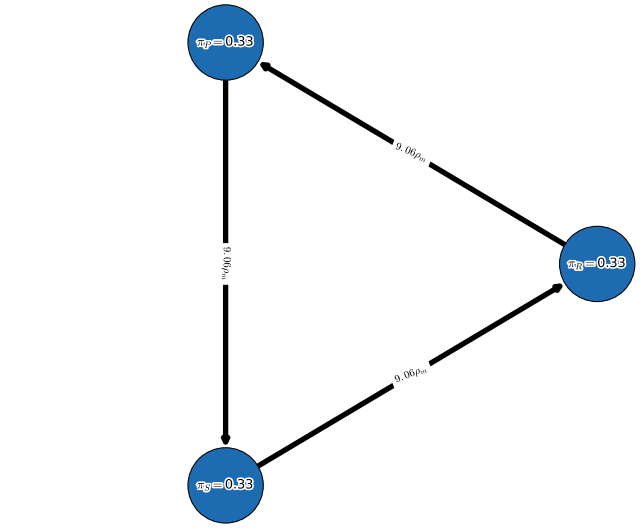
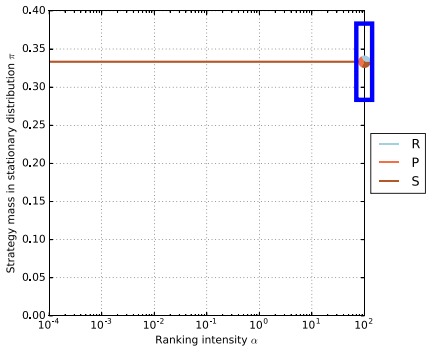


Figure 5. Topology of dynamical systems: an (ε, T)-chain from x_0 to x_4 with respect to flow φ is exemplified here by the solid arrows and sequence of points x_0, x_1, x_2, x_3, x_4. If the recurrent behavior associated with point x_0 (indicated by the dashed arrow) holds for all $\varepsilon > 0$ and $T > 0$, then it is a chain recurrent point.



(a)



(b)

Agent	Rank	Score
R	1	0.33
P	1	0.33
S	1	0.33

(c)

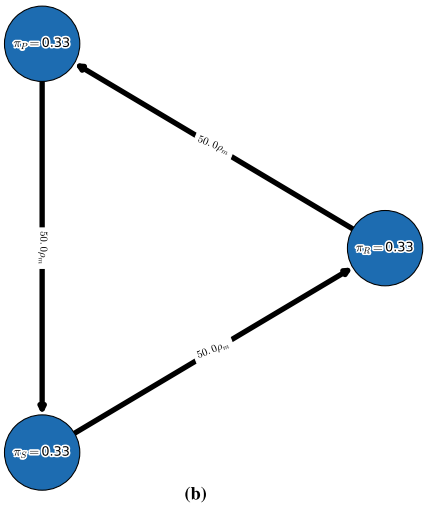
Figure 6. Rock-Paper-Scissors game. (**a**) Discrete-time dynamics. (**b**) Ranking-intensity sweep. (**c**) α-Rank results.

$$\dot{x}_i = x_i((Ay)_i - x^T Ay) \quad \dot{y}_j = y_j((x^T B)_j - x^T By) \quad \forall\ (i,j) \in S^1 \times S^2, \tag{1}$$

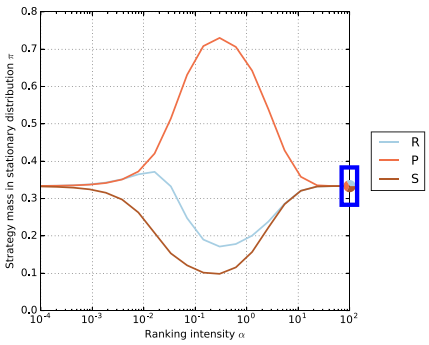
where x_i and y_j are, respectively, the proportions of strategies $i \in S^1$ and $j \in S^2$ in two infinitely-sized populations, each corresponding to a player. This system of coupled differential equations models the temporal dynamics of the populations' strategy profiles when they interact, and can be extended readily to the general K-wise interaction case (see Supplementary Material Appendix S2.1 for more details).

		Player 2		
		R	*P*	*S*
Player 1	*R*	0	−0.5	1
	P	0.5	0	−0.1
	S	−1	0.1	0

(a)



(b)



(c)

Agent	Rank	Score
R	1	0.33
P	1	0.33
S	1	0.33

(d)

Figure 7. Biased Rock-Paper-Scissors game. (**a**) Payoff matrix. (**b**) Discrete-time dynamics. (**c**) Ranking-intensity sweep. (**d**) α-Rank results.

The replicator dynamics provide useful insights into the micro-dynamical characteristics of games, revealing strategy flows, basins of attraction, and equilibria[36] when visualized on a trajectory plot over the strategy simplex (e.g., Fig. 4). The accessibility of these insights, however, becomes limited for games involving large strategy spaces and many-player interactions. For instance, trajectory plots may be visualized only for subsets of three or four strategies in a game, and are complex to analyze for multi-population games due to the inherently-coupled nature of the trajectories. While methods for scalable empirical game-theoretic analysis of games have been recently introduced, they are still limited to two-population games[6,7].

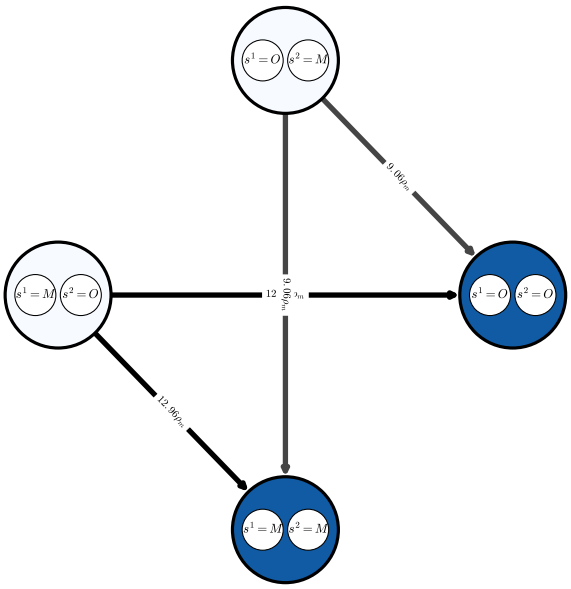
Macro-model: discrete-time dynamics. This section presents our main evolutionary dynamics model, which extends previous single-population discrete-time models and is later shown to play an integral role in our α-Rank method and can also be seen as an instantiation of the framework introduced in[20].

A promising alternative to using the continuous-time replicator dynamics for evaluation is to consider discrete-time finite-population dynamics. As later demonstrated, an important advantage of the discrete-time dynamics is that they are not limited to only three or four strategies (i.e., the agents under evaluation) as in the continuous-time case. Even though we lose the micro-dynamical details of the strategy simplex, this discrete-time macro-dynamical model, in which we observe the flows over the edges of the high-dimensional simplex, still provides useful insights into the overall system dynamics.

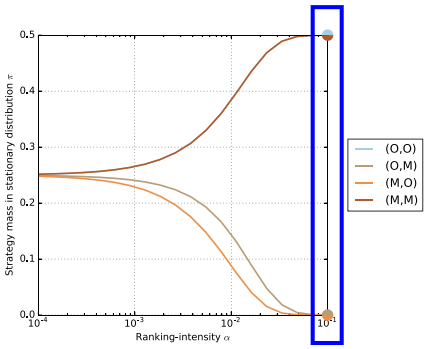
To conduct this discrete-time analysis, we consider a selection-mutation process but with a *very small mutation rate* (following the small mutation rate theorem, see[37]). Before elaborating on the details we specify a number of important concepts used in the description below and clarify their respective meanings in Fig. 2a. Let a *monomorphic population* denote a population wherein all individuals play identical strategies, and a *monomorphic*

Domain	Results	Symmetric?	# of Populations	# of Strategies
Rock-Paper-Scissors	Section 3.2.1	✓	1	[3]
Biased Rock-Paper-Scissors	Section 3.2.2	✓	1	[3]
Battle of the Sexes	Section 3.2.3	✗	2	[2, 2]
AlphaGo	Section 3.4.1	✓	1	[7]
AlphaZero Chess	Section 3.4.2	✓	1	[56]
MuJoCo Soccer	Section 3.4.3	✓	1	[10]
Kuhn Poker	Section 3.4.4	✗	3	[4, 4, 4]
	Section 3.4.4	✗	4	[4, 4, 4, 4]
Leduc Poker	Section 3.4.5	✗	2	[3, 3]

Table 1. Overview of multi-agent domains evaluated in this paper. These domains are extensive across multiple axes of complexity, and include symmetric and asymmetric games with different numbers of populations and ranges of strategies.



(a)



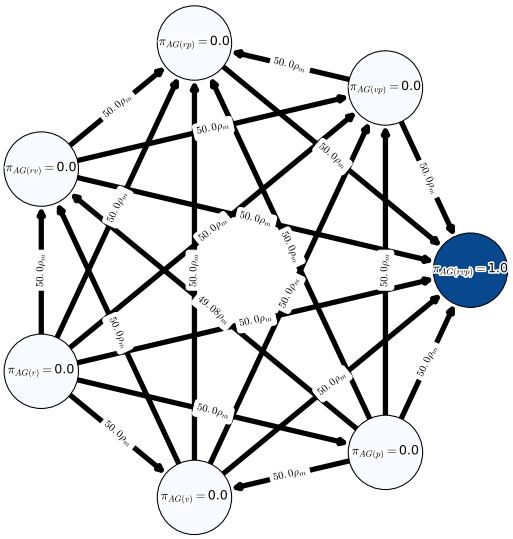
(b)

Agent	Rank	Score
(O,O)	1	0.5
(M,M)	1	0.5
(O,M)	2	0.0
(M,O)	2	0.0

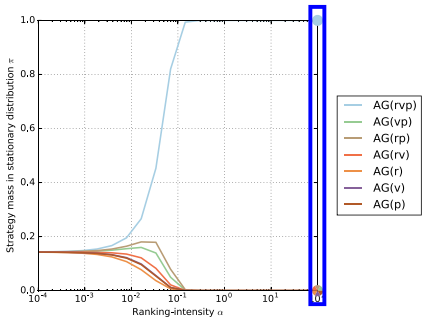
(c)

Figure 8. Battle of the Sexes. (**a**) Discrete-time dynamics (see (**c**) for node-wise scores corresponding to stationary distribution masses). (**b**) Ranking-intensity sweep. (**c**) α-Rank results.

population profile denote a set of monomorphic populations, where each population may be playing a different strategy (see Fig. 2b). Our general idea is to capture the overall dynamics by defining a *Markov chain* over states that correspond to monomorphic population profiles. We can then calculate the transition probability matrix over these states, which captures the fixation probability of any mutation in any given population



(a)



(b)

Agent	Rank	Score
$AG(rvp)$	1	1.0
$AG(vp)$	2	0.0
$AG(rp)$	2	0.0
$AG(rv)$	2	0.0
$AG(r)$	2	0.0
$AG(v)$	2	0.0
$AG(p)$	2	0.0

(c)

Figure 9. AlphaGo (Nature dataset). (**a**) Discrete-time dynamics. (**b**) Ranking-intensity sweep. (**c**) α-Rank results.

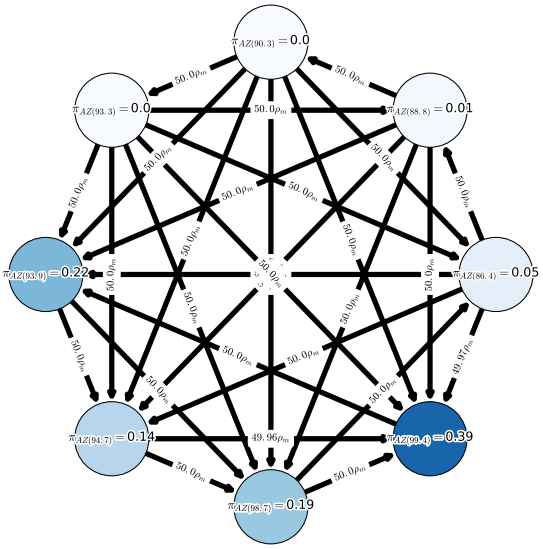
(i.e., the probability that the mutant will take over that population). By computing the stationary distribution over this matrix we find the evolutionary population dynamics, which can be represented as a graph. The nodes of this graph correspond to the states, with the stationary distribution quantifying the average time spent by the populations in each node[19,38].

A large body of prior literature has conducted this discrete-time Markov chain analysis in the context of pair-wise interaction games with symmetric payoffs[16,17,19,38,39]. Recent work applies the underlying assumption of small-mutation rates[37] to propose a general framework for discrete-time multi-player interactions[20], which applies to games with asymmetric payoffs. In our work, we formalize how such an evolutionary model, in the micro/macro dynamics spectrum, should be instantiated to converge to our novel and dynamical solution concept of MCCs. Additionally, we show (in Theorem 2.1.3) that in the case of identical per-population payoffs (i.e., $\forall\ k,\ M^k = M$) our generalization reduces to the single-population model used by prior works. For completeness, we also detail the single population model in the Supplementary Material (see Appendix S2.2).We now formally define the generalized discrete-time model.

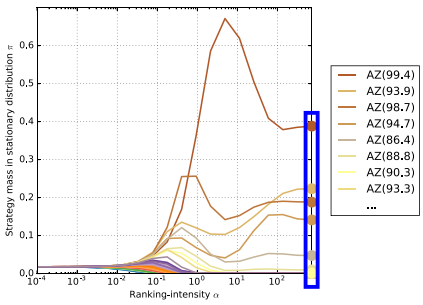
Recall from Section 2.1.1 that each individual in a K-wise interaction game receives a local payoff $M^k(s^k, s^{-k})$ dependent on its identity k, its strategy s^k, and the strategy profile s^{-k} of the other $K-1$ individuals involved in the game. To account for the identity-dependent payoffs of such individuals, we consider the interactions of K finite populations, each corresponding to a specific identity $k \in \{1, \ldots, K\}$.

In each population k, we have a set of strategies S^k that we would like to evaluate for their evolutionary strength. We also have a set of individuals A in each population k, each of which is programmed to play a strategy from the set S^k. Without loss of generality, we assume all populations have m individuals.

Individuals interact K-wise through empirical games. At each timestep T, one individual from each population is sampled uniformly, and the K resulting individuals play a game. Let $p^k_{s^k}$ denote the number of individuals in population k playing strategy s^k and p denote the joint population state (i.e., vector of states of all populations). Under our sampling protocol, the fitness of an individual that plays strategy s^k is,



(a)



(b)

Agent	Rank	Score
$AZ(99.4)$	1	0.39
$AZ(93.9)$	2	0.22
$AZ(98.7)$	3	0.19
$AZ(94.7)$	4	0.14
$AZ(86.4)$	5	0.05
$AZ(88.8)$	6	0.01
$AZ(90.3)$	7	0.0
$AZ(93.3)$	8	0.0
...	...	...

(c)

Figure 10. AlphaZero dataset. (**a**) Discrete-time dynamics. (**b**) Ranking-intensity sweep. (**c**) α-Rank results.

$$f^k(s^k, p^{-k}) = \sum_{s^{-k} \in S^{-k}} M^k(s^k, s^{-k}) \prod_{c \in \{1,\ldots,K\} \setminus k} \frac{p^c_{s^c}}{m}. \tag{2}$$

We consider any two individuals from a population k, with respective strategies τ, $\sigma \in S^k$ and respective fitnesses $f^k(\tau, p^{-k})$ and $f^k(\sigma, p^{-k})$ (dependent on the values of the meta-game table). We introduce here a discrete-time dynamics, where the strategy of the first individual (playing τ) is then updated by either mutating with a very small probability to a random strategy (Fig. 2c), probabilistically copying the strategy σ of the second individual (Fig. 2d), or sticking with its own strategy τ. The idea is that strong individuals will replicate and spread throughout the population (Fig. 2e). While one could choose other variants of discrete-time dynamics[40], we show that this particular choice both yields useful closed-form representations of the limiting behaviors of the populations, and also coincides with the MCC solution concept we later introduce under specific conditions.

As individuals from the same population never directly interact, the state of a population k has no bearing on the fitnesses of its individuals. However, as evident in (2), each population's fitness may directly be affected by the competing populations' states. The complexity of analyzing such a system can be significantly reduced by making the assumption of a small mutation rate[37]. Let the 'focal population' denote a population k wherein a mutant strategy appears. We denote the probability for a strategy to mutate randomly into another strategy $s^k \in S^k$ by μ and we will assume it to be infinitesimally small (i.e., we consider a small-mutation limit $\mu \to 0$). If we neglected mutations, the end state of this evolutionary process would be monomorphic. If we introduce a very small mutation rate this means that either the mutant fixates and takes over the current population, or the current population is capable of wiping out the mutant strategy[37]. Therefore, given a small mutation rate, the mutant almost always either fixates or disappears before a new mutant appears in the current population. This means that any given population k will almost never contain more than two strategies at any point in time. We refer the interested reader to[20] for a more extensive treatment of these arguments.

Applying the same line of reasoning, in the small-mutation rate regime, the mutant strategy in the focal population will either fixate or go extinct much earlier than the appearance of a mutant in any *other* population[37]. Thus, at any given time, there can maximally be only one population with a mutant, and the remaining

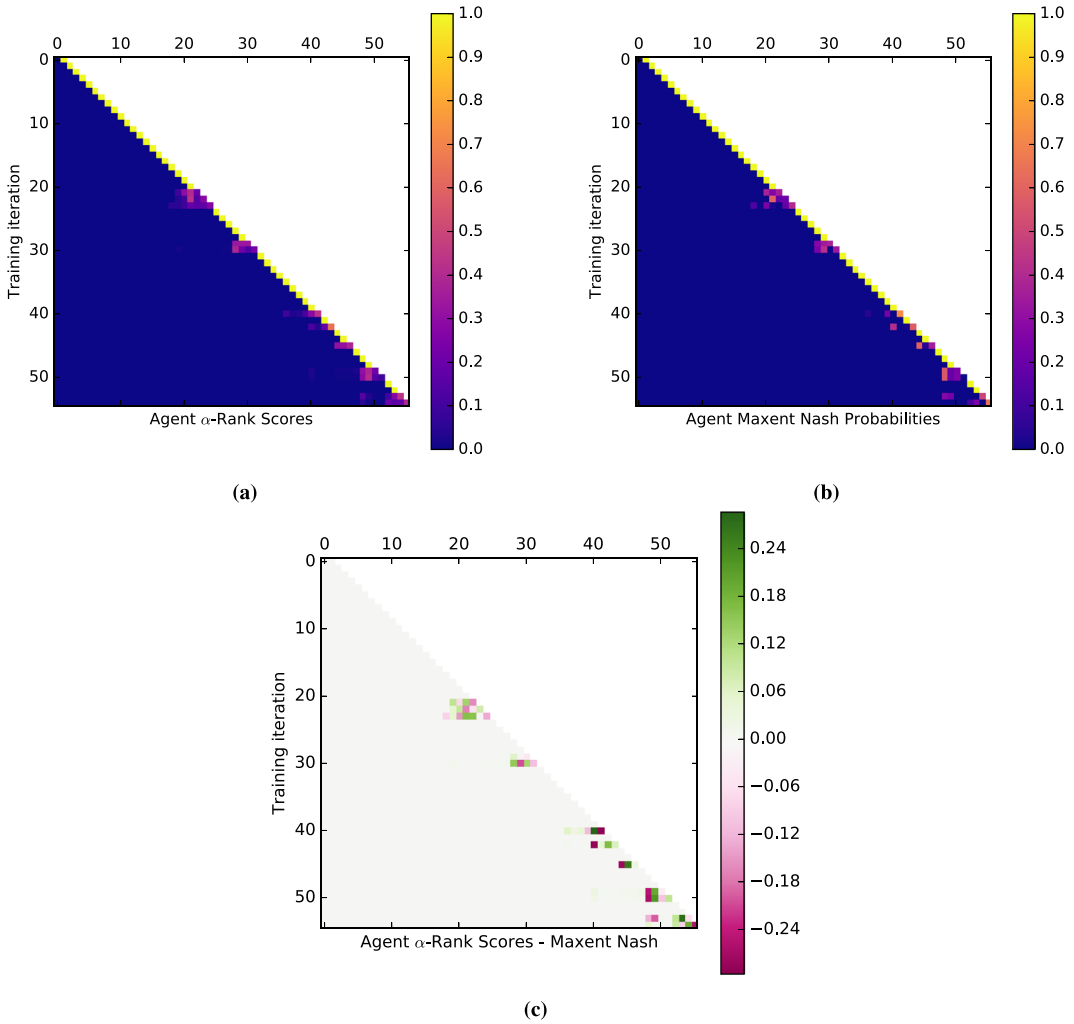


Figure 11. AlphaZero (chess) agent evaluations throughout training. (**a**) α-Score vs. Training Time. (**b**) Maximum Entropy Nash vs. Training Time. (**c**) α-Score - Maximum Entropy Nash difference.

populations will be monomorphic; i.e., in each competing population $c \in \{1, \ldots, K\} \backslash k$, $\frac{p^c_{s^c}}{m} = 1$ for a single strategy and 0 for the rest. As such, given a small enough mutation rate, analysis of any focal population k needs only consider the monomorphic states of all other populations. Overloading the notation in (2), the fitness of an individual from population k that plays s^k then considerably simplifies to

$$f^k(s^k, s^{-k}) = M^k(s^k, s^{-k}), \tag{3}$$

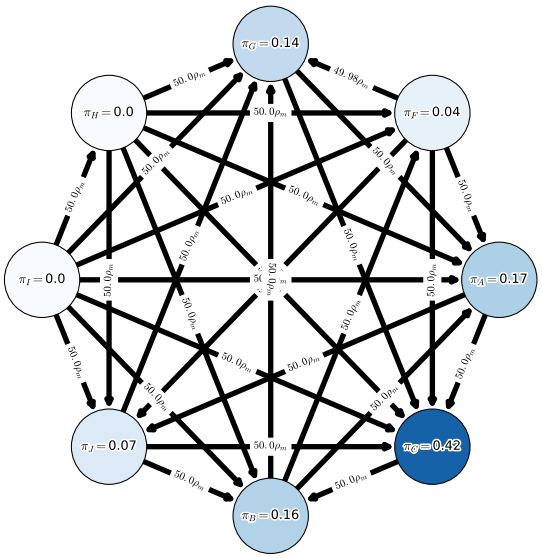
where s^{-k} denotes the strategy profile of the other populations.

Let p^k_τ and p^k_σ respectively denote the number of individuals playing τ and σ in focal population k, where $p^k_\tau + p^k_\sigma = m$. Per (3), the fitness of an individual playing τ in the focal population while the remaining populations play monomorphic strategies s^{-k} is given by $f^k(\tau, s^{-k}) = M^k(\tau, s^{-k})$. Likewise, the fitness of any individual in k playing σ is, $f^k(\sigma, s^{-k}) = M^k(\sigma, s^{-k})$.

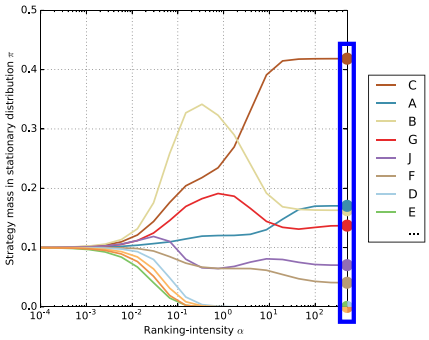
We randomly sample two individuals in population k and consider the probability that the one playing τ copies the other individual's strategy σ. The probability with which the individual playing strategy τ will copy the individual playing strategy σ can be described by a *selection* function $\mathbb{P}(\tau \to \sigma, s^{-k})$, which governs the dynamics of the finite-population model. For the remainder of the paper, we focus on the logistic selection function (aka Fermi distribution),

$$\mathbb{P}(\tau \to \sigma, s^{-k}) = \frac{e^{\alpha f^k(\sigma, s^{-k})}}{e^{\alpha f^k(\tau, s^{-k})} + e^{\alpha f^k(\sigma, s^{-k})}} = \left(1 + e^{\alpha(f^k(\tau, s^{-k}) - f^k(\sigma, s^{-k}))}\right)^{-1}, \tag{4}$$

with α determining the selection strength, which we call the *ranking-intensity* (the correspondence between α and our ranking method will become clear later). There are alternative definitions of the selection function that



(a)



(b)

Agent	Rank	Score
C	1	0.42
A	2	0.17
B	3	0.16
G	4	0.14
J	5	0.07
F	6	0.04
D	7	0.0
E	7	0.0
...	...	...

(c)

Figure 12. MuJoCo soccer dataset. (**a**) Discrete-time dynamics. (**b**) Ranking-intensity sweep. (**c**) α-Rank results.

may be used here, we merely focus on the Fermi distribution due to its extensive use in the single-population literature[16,17,19].

Based on this setup, we define a Markov chain over the set of strategy profiles $\prod_k S^k$ with $\prod_k |S^k|$ states. Each state corresponds to one of the strategy profiles $s \in \prod_k S^k$, representing a multi-population end-state where each population is monomorphic. The transitions between these states are defined by the corresponding fixation probabilities (the probability of overtaking the population) when a mutant strategy is introduced in any single monomorphic population k. We now define the Markov chain, which has $(\prod_k |S^k|)^2$ transition probabilities over all pairs of monomorphic multi-population states. Denote by $\rho^k_{\sigma,\tau}(s^{-k})$ the probability of mutant strategy τ fixating in a focal population k of individuals playing σ, while the remaining $K-1$ populations remain in their monomorphic states s^{-k}. For any given monomorphic strategy profile, there are a total of $\sum_k(|S^k|-1)$ valid transitions to a subsequent profile where only a single population has changed its strategy. Thus, letting $\eta = \frac{1}{\sum_k(|S^k|-1)}$, then $\eta\rho^k_{\sigma,\tau}(s^{-k})$ is the probability that the joint population state transitions from (σ, s^{-k}) to state (τ, s^{-k}) after the occurrence of a single mutation in population k. The stationary distribution over this Markov chain tells us how much time, on average, the dynamics will spend in each of the monomorphic states.

The fixation probabilities (of a rare mutant playing τ overtaking the focal population k) can be calculated as follows. The probability that the number of individuals playing τ decreases/increases by one in the focal population is given by,

$$T^{k(\mp 1)}(p^k, \tau, \sigma, s^{-k}) = \frac{p^k_\tau p^k_\sigma}{m(m-1)}\left(1 + e^{\pm\alpha(f^k(\tau,s^{-k})-f^k(\sigma,s^{-k}))}\right)^{-1}. \tag{5}$$

The fixation probability $\rho^k_{\sigma,\tau}(s^{-k})$ of a single mutant with strategy τ in a population k of $m-1$ individuals playing σ is derived as follows. Let $u = f^k(\tau, s^{-k}) - f^k(\sigma, s^{-k})$, then,

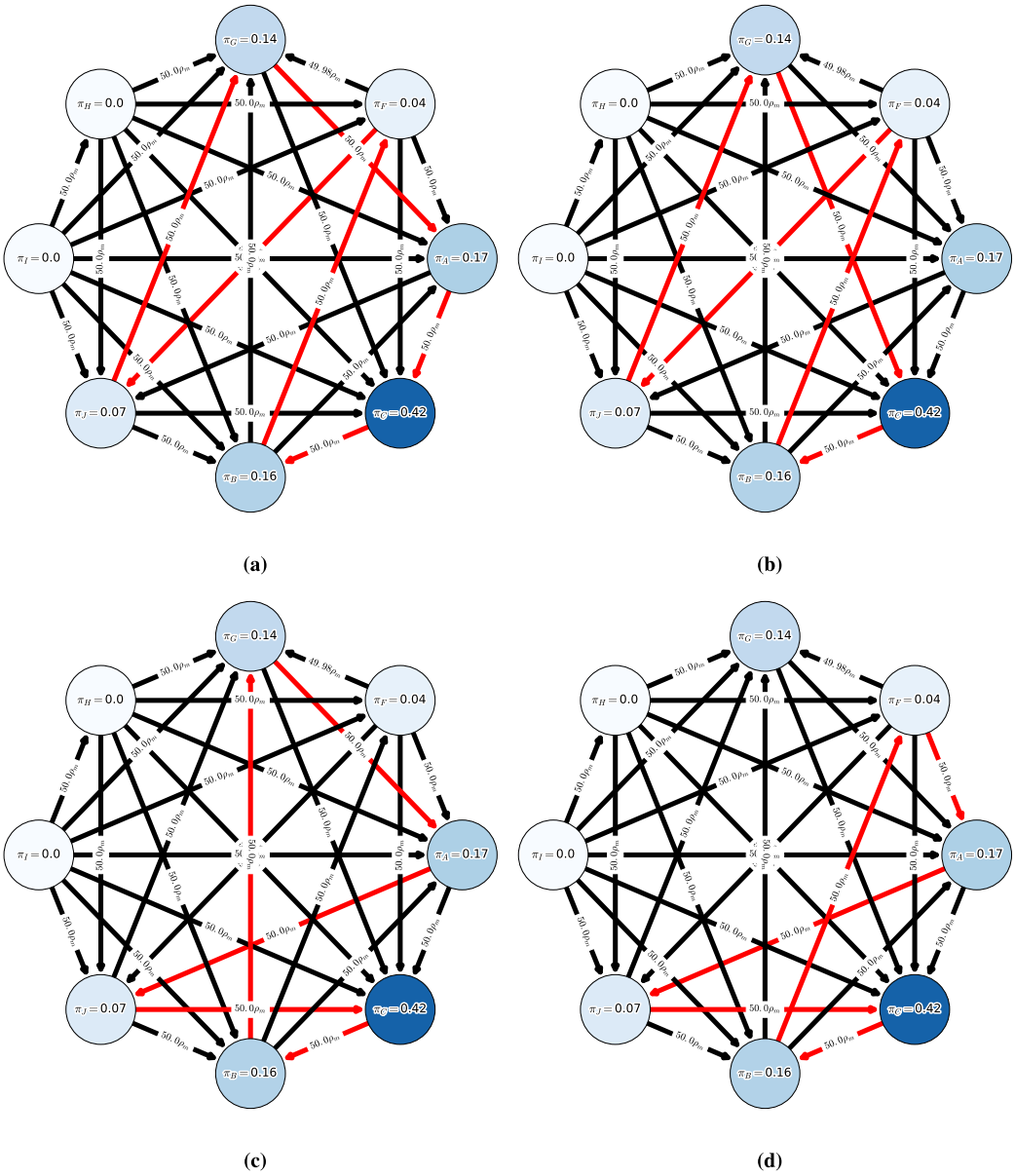
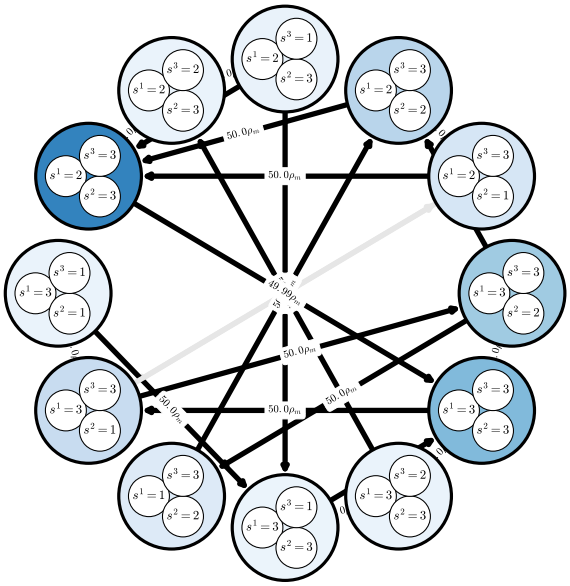


Figure 13. Example cycles in the MuJoCo soccer domain.

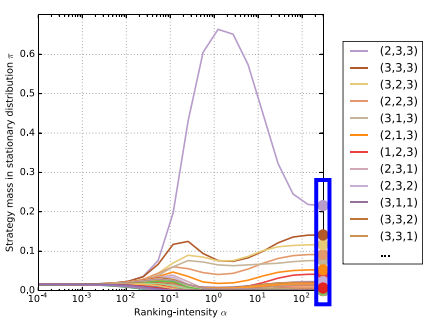
$$\rho^k_{\sigma,\tau}(s^{-k}) = \left(1 + \sum_{l=1}^{m-1} \prod_{p^k_\tau=1}^{l} \frac{T^{k(-1)}(p^k, \tau, \sigma, s^{-k})}{T^{k(+1)}(p^k, \tau, \sigma, s^{-k})}\right)^{-1} \tag{6}$$

$$= \left(1 + \sum_{l=1}^{m-1} \prod_{p^k_\tau=1}^{l} \frac{(1 + e^{\alpha u})^{-1}}{(1 + e^{-\alpha u})^{-1}}\right)^{-1} \tag{7}$$

$$= \left(1 + \sum_{l=1}^{m-1} \prod_{p^k_\tau=1}^{l} \frac{1 + e^{-\alpha u}}{1 + e^{\alpha u}}\right)^{-1} \tag{8}$$



(a)



(b)

Agent	Rank	Score
(2,3,3)	1	0.22
(3,3,3)	2	0.14
(3,2,3)	3	0.12
(2,2,3)	4	0.09
(3,1,3)	5	0.08
(2,1,3)	6	0.05
(1,2,3)	7	0.04
(2,3,1)	8	0.02
(2,3,2)	9	0.02
(3,1,1)	10	0.02
(3,3,2)	11	0.02
(3,3,1)	12	0.02
...	...	...

(c)

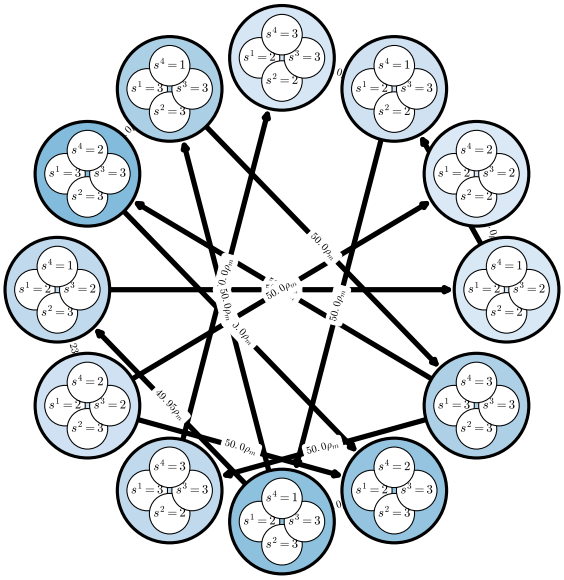
Figure 14. 3-player Kuhn poker (ranking conducted on all 64 pure strategy profiles). (**a**) Discrete-time dynamics. (**b**) Ranking-intensity sweep. (**c**) α-Rank results.

$$= \left(1 + \sum_{l=1}^{m-1} \prod_{p_\tau^k=1}^{l} \frac{\frac{e^{\alpha u}+1}{e^{\alpha u}}}{1+e^{\alpha u}}\right)^{-1} \tag{9}$$

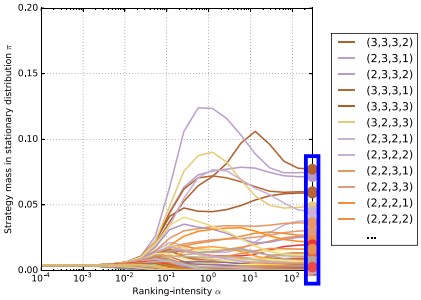
$$= \left(1 + \sum_{l=1}^{m-1} \prod_{p_\tau^k=1}^{l} e^{-\alpha u}\right)^{-1} \tag{10}$$

$$= \left(1 + \sum_{l=1}^{m-1} \prod_{p_\tau^k=1}^{l} e^{-\alpha(f^k(\tau,s^{-k})-f^k(\sigma,s^{-k}))}\right)^{-1} \tag{11}$$

$$= \left(1 + \sum_{l=1}^{m-1} e^{-l\alpha(f^k(\tau,s^{-k})-f^k(\sigma,s^{-k}))}\right)^{-1} \tag{12}$$



(a)



(b)

Agent	Rank	Score
(3,3,3,2)	1	0.08
(2,3,3,1)	2	0.07
(2,3,3,2)	3	0.07
(3,3,3,1)	4	0.06
(3,3,3,3)	5	0.06
(3,2,3,3)	6	0.05
(2,3,2,1)	7	0.04
(2,3,2,2)	8	0.04
(2,2,3,1)	9	0.04
(2,2,3,3)	10	0.03
(2,2,2,1)	11	0.03
(2,2,2,2)	12	0.03
...	...	...

(c)

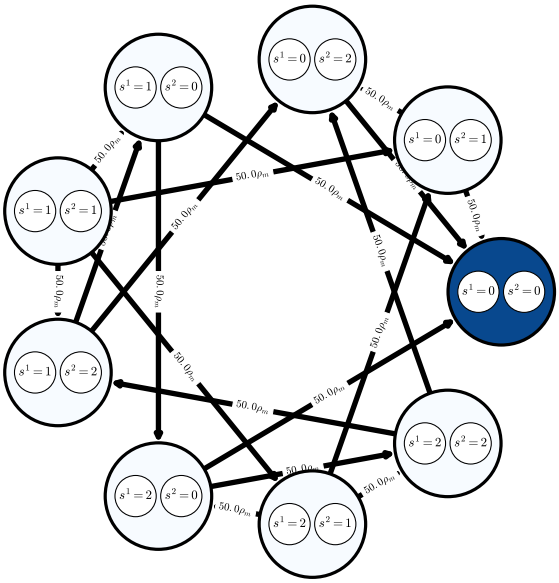
Figure 15. 4-player Kuhn poker (ranking conducted on all 256 pure strategy profiles). (**a**) Discrete-time dynamics. (**b**) Ranking-intensity sweep. (**c**) α-Rank results.

$$= \begin{cases} \dfrac{1 - e^{-\alpha(f^k(\tau,s^{-k}) - f^k(\sigma,s^{-k}))}}{1 - e^{-m\alpha(f^k(\tau,s^{-k}) - f^k(\sigma,s^{-k}))}} & \text{if } f^k(\tau, s^{-k}) \neq f^k(\sigma, s^{-k}) \\ \dfrac{1}{m} & \text{if } f^k(\tau, s^{-k}) = f^k(\sigma, s^{-k}) \end{cases} \tag{13}$$

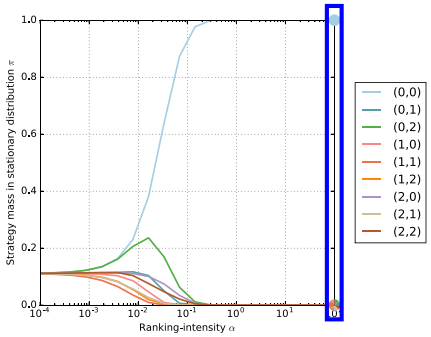
This corresponds to the computation of an m-step transition in the Markov chain corresponding to $\mathbb{P}(\tau \to \sigma, s^{-k})$[41]. The quotient $\frac{T^{k(-1)}(p^k,\tau,\sigma,s^{-k})}{T^{k(+1)}(p^k,\tau,\sigma,s^{-k})}$ expresses the likelihood (odds) that the mutation process in population k continues in either direction: if it is close to zero then it is very likely that the number of mutants (individuals with strategy τ in population k) increases; if it is very large it is very likely that the number of mutants will decrease; and if it close to one then the probabilities of increase and decrease of the number of mutants are equally likely. This yields the following Markov transition matrix corresponding to the jump from strategy profile $s_i \in \prod_k S^k$ to $s_j \in \prod_k S^k$,

$$C_{ij} = \begin{cases} \eta \rho^{\,k}_{s_i^k, s_j^k}(s_i^{-k}) & \text{if } \exists k \text{ such that } s_i^k \neq s_j^k \text{ and } s_i^{-k} = s_j^{-k} \\ 1 - \sum_{j \neq i} C_{ij} & \text{if } s_i = s_j \\ 0 & \text{otherwise} \end{cases} \tag{14}$$

for all $i, j \in \{1, \ldots, |S|\}$, where $|S| = \prod_k |S^k|$.



(a)



(b)

Agent	Rank	Score
(0,0)	1	1.0
(0,1)	2	0.0
(0,2)	2	0.0
(1,0)	2	0.0
(1,1)	2	0.0
(1,2)	2	0.0
(2,0)	2	0.0
(2,1)	2	0.0
(2,2)	2	0.0

(c)

Figure 16. PSRO poker dataset. (**a**) Discrete-time dynamics (top 8 agents shown only). (**b**) Ranking-intensity sweep. (**c**) α-Rank strategy rankings and scores (top 8 agents shown only).

The following theorem formalizes the irreducibility of this finite-population Markov chain, a property that is well-known in the literature (e.g., see [37, Theorem 2] and [20, Theorem 1]) but stated here for our specialized model for completeness.

Theorem 2.1.2 Given finite payoffs, the Markov chain with transition matrix C is irreducible (i.e., it is possible to get to any state starting from any state). Thus a unique stationary distribution π (where $\pi^T C = \pi^T$ and $\sum_i \pi_i = 1$) exists.

Proof. Refer to the Supplementary Material for the proof. □

This unique π provides the evolutionary ranking, or strength of each strategy profile in the set $\prod_k S^k$, expressed as the average time spent in each state in distribution π.

This generalized discrete-time evolutionary model, as later shown, will form the basis of our α-Rank method. We would like to clarify the application of this general model to the single population case, which applies only to symmetric 2-player games and is commonly used in the literature (see Appendix S1).

Application to Single-Population (Symmetric Two-Player) Games. For completeness, we provide a detailed outline of the single population model in the Supplementary Material Appendix S2.2. We also include remarks regarding the validity of the monomorphic population assumption, as used in our model and those of prior works.

Theorem 2.1.3 (Multi-population model generalizes the symmetric single-population model). The general multi-population model inherently captures the dynamics of the single population symmetric model.

Proof. (Sketch) In the pairwise symmetric game setting, we consider only a single population of interacting individuals (i.e., $K = 1$), where a maximum of two strategies may exist at any time in the population due to the small mutation rate assumption. At each timestep, two individuals (with respective strategies τ, $\sigma \in S^1$) are sampled from this population and play a game using their respective strategies τ and σ. Their respective fitnesses then

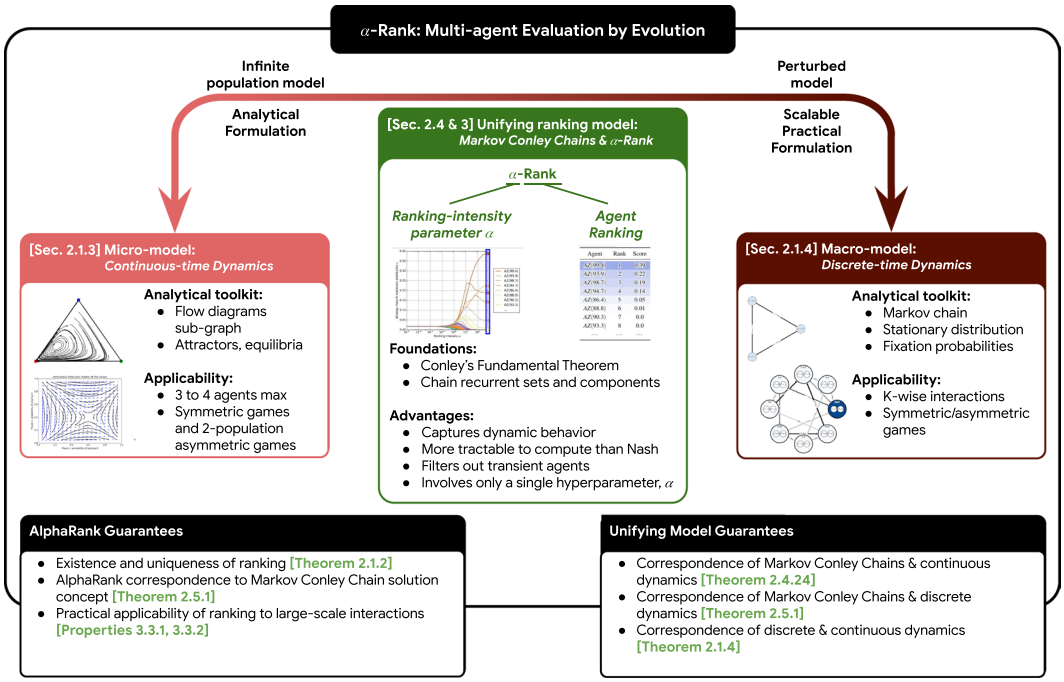


Figure 17. A retrospective look on the paper contributions. We introduced a general descriptive multi-agent evaluation method, called α-*Rank*, which is practical in the sense that it is easily applicable in complex game-theoretic settings, and theoretically-grounded in a solution concept called Markov-Conley chains (MCCs). α-Rank has a strong theoretical and specifically evolutionary interpretation; the overarching perspective considers a chain of models of increasing complexity, with a discrete-time macro-dynamics model on one end, continuous-time micro-dynamics on the other end, and MCCs as the link in between. We provided both scalability properties and theoretical guarantees for the overall ranking methodology.

correspond directly to their payoffs, i.e., $f_\tau = M(\tau, \sigma)$ and $f_\sigma = M(\sigma, \tau)$. With this change, all other derivations and results follow directly the generalized model. For example, the probability of decrease/increase of a strategy of type s_τ in the single-population case translates to,

$$T^{(\mp 1)}(p, \tau, \sigma) = \frac{p_\tau p_\sigma}{m(m-1)}(1 + e^{\pm\alpha(f_\tau - f_\sigma)})^{-1}, \tag{15}$$

and likewise for the remaining equations. □

In other words, the generalized model is general in the sense that one can not only simulate symmetric pairwise interaction dynamics, but also K-wise and asymmetric interactions.

Linking the Micro- and Macro-dynamics Models. We have introduced, so far, a micro- and macro-dynamics model, each with unique advantages in terms of analyzing the evolutionary strengths of agents. The formal relationship between these two models remains of interest, and is established in the limit of a large population:

Theorem 2.1.4 (Discrete-Continuous Edge Dynamics Correspondence). In the large-population limit, the macro-dynamics model is equivalent to the micro-dynamics model over the edges of the strategy simplex. Specifically, the limiting model is a variant of the replicator dynamics with the caveat that the Fermi revision function takes the place of the usual fitness terms.

Proof. Refer to the Supplementary Material for the proof. □

Therefore, a correspondence exists between the two models on the 'skeleton' of the simplex, with the macro-dynamics model useful for analyzing the global evolutionary behaviors over this skeleton, and the micro-model useful for 'zooming into' the three- or four-faces of the simplex to analyze the interior dynamics.

In the next sections, we first give a few conceptual examples of the generalized discrete-time model, then discuss the need for a new solution concept and the incompatibility between Nash equilibria and dynamical systems. We then directly link the generalized model to our new game-theoretic solution concept, Markov-Conley chains (in Theorem 2.5.1).

Conceptual examples. We present two canonical examples that visualize the discrete-time dynamics and build intuition regarding the macro-level insights gained using this type of analysis.

Rock-Paper-Scissors. We first consider the single-population (symmetric) discrete-time model in the Rock-Paper-Scissors (RPS) game, with the payoff matrix shown in Fig. 3a (top). One can visualize the discrete-time dynamics using a graph that corresponds to the Markov transition matrix C defined in (14), as shown in Fig. 3a (bottom).

Nodes in this graph correspond to the monomorphic population states. In this example, these are the states of the population where all individuals play as agents Rock, Paper, or Scissors. To quantify the time the population spends as each agent, we indicate the corresponding mass of the stationary distribution π within each node. As can be observed in the graph, the RPS population spends exactly $\frac{1}{3}$ of its time as each agent.

Edges in the graph correspond to the fixation probabilities for pairs of states. Edge directions corresponds to the flow of individuals from one agent to another, with strong edges indicating rapid flows towards 'fitter' agents. We denote fixation probabilities as a multiple of the neutral fixation probability baseline, $\rho_m = \frac{1}{m}$, which corresponds to using the Fermi selection function with $\alpha = 0$. To improve readability of the graphs, we also do not visualize edges looping a node back to itself, or edges with fixation probabilities lower than ρ_m. In this example, we observe a cycle (intransitivity) involving all three agents in the graph. While for small games such cycles may be apparent directly from the structure of the payoff table, we later show that the graph visualization can be used to automatically iterate through cycles even in K-player games involving many agents.

Battle of the sexes. Next we illustrate the generalized multi-population (asymmetric) model in the Battle of the Sexes game, with the payoff matrix shown in Fig. 3b (top). The graph now corresponds to the interaction of two populations, each representing a player type, with each node corresponding to a monomorphic population *profile* (s^1, s^2). Edges, again, correspond to fixation probabilities, but occur only when a single population changes its strategy to a different one (an artifact of our small mutation assumption). In this example, it is evident from the stationary distribution that the populations spend an equal amount of time in profiles (O, O) and (M, M), and essentially zero time in states (O, M) and (M, O).

The incompatibility of nash equilibrium and dynamical systems. Continuous- and discrete-time dynamical systems have been used extensively in Game Theory, Economics, and Algorithmic Game Theory. In the particular case of multi-agent evaluation in meta-games, this type of analysis is relied upon for revealing useful insights into the strengths and weaknesses of interacting agents[6]. Often, the goal of research in these areas is to establish that, in some sense, the investigated dynamics actually converge to a Nash equilibrium; there has been limited success in this front, and there are some negative results[42–44]. In fact, all known dynamics in games (the replicator dynamics, the many continuous variants of the dynamics used in the proof of Nash's theorem, etc.) do cycle. To compound this issue, meta-games are often large, extend beyond pair-wise interactions, and may not be zero-sum. While solving for a Nash equilibrium can be done in polynomial time for zero-sum games, doing so in general-sum games is known to be PPAD-complete[21], which severely limits the feasibility of using such a solution concept for evaluating our agents.

Of course, some dynamics are known to converge to *relaxations* of the Nash equilibrium, such as the correlated equilibrium polytope or the coarse correlated equilibria[45]. Unfortunately, this "convergence" is typically considered in the sense of *time average*; time averages can be useful for establishing performance bounds for games, but tell us little about actual system behavior — which is a core component of what we study through games. For certain games, dynamics may indeed converge to a Nash equilibrium, but they may also *cycle*. For example, it is encouraging that in all 2×2 matrix games these equilibria, cycles, and slight generalizations thereof are the only possible limiting behaviors for continuous-time dynamics (i.e., flows). Unfortunately, this clean behavior (convergence to either a cycle or, as a special case, to a Nash equilibrium) is an artifact of the two-dimensional nature of 2×2 games, a consequence of the Poincaré–Bendixson theorem[46]. There is a wide range of results in different disciplines arguing that learning dynamics in games tend to not equilibrate to any Nash equilibrium but instead exhibit complex, unpredictable behavior (e.g.[42,47–51]). The dynamics of even simple two-person games with three or more strategies per player can be chaotic[52], that is, inherently difficult to predict and complex. Chaos goes against the core of our objectives, leaving little hope for building a predictive theory of player behavior based on dynamics in terms of Nash equilibrium.

Markov-Conley chains: A dynamical solution concept. Recall our overall objective: we would like to understand and evaluate multi-agent interactions using a detailed and realistic model of evolution, such as the replicator dynamics, in combination with a game-theoretic solution concept. We start by acknowledging the fundamental incompatibility between dynamics and the Nash equilibrium: dynamics are often incapable of reaching the Nash equilibrium. However, instead of taking this as a disappointing flaw of dynamics, we see it instead as an opportunity to look for a novel solution concept that does not have the same limitations as Nash in relation to these dynamical systems. We contemplate whether a plausible algorithmic solution concept can emerge by asking, *what do these dynamics converge to?* Our goal is to identify the non-trivial, irreducible behaviors of a dynamical system and thus provide a new solution concept — an alternative to Nash's — that will enable evaluation of of multi-agent interactions using the underlying evolutionary dynamics. We carve a pathway towards this alternate solution concept by first considering the topology of dynamical systems.

Topology of dynamical systems and conley's theorem. Dynamicists and topologists have seeked means of extending to higher dimensions the benign, yet complete, limiting dynamical behaviors described in Section 2.3 that one sees in two dimensions: convergence to cycles (or equilibria as a special case). That is, they have been trying to find an appropriate *relaxation of the notion of a cycle* such that the two-dimensional picture is restored. New conceptions of "periodicity" and "cycles" were indeed discovered, in the form of *chain recurrent sets* and *chain components*, which we define in this section. These key ingredients form the foundation of Conley's Fundamental Theorem of Dynamical Systems, which in turn leads to the formulation of our Markov-Conley chain solution concept and associated multi-agent evaluation scheme.

Definitions. To make our contribution formal, we need define certain topological concepts, following the treatment of Conley[29]. Our chain recurrence approach and the theorems in this section follow from[53]. We also provide the interested reader a general background on dynamical systems in Supplementary Material S2 in order to make our work self-contained.

Let ϕ: $\mathbb{R} \times X \rightarrow$ denote a flow on a topological space X. We sometimes write $\phi^t(x)$ for $\phi(t, x)$ and denote a flow ϕ: $\mathbb{R} \times X \rightarrow X$ by ϕ^t: $X \rightarrow X$, where $t \in \mathbb{R}$. For more background on dynamical systems see section S2 in the appendix.

Definition 2.4.1 ((ε, T)-chain). Let ϕ be a flow on a metric space (X, d). Given $\varepsilon > 0, T > 0$, and $x, y \in X$, an (ε, T)-chain from x to y with respect to ϕ and d is a pair of finite sequences $x = x_0, x_1, \ldots, x_{n-1}, x_n = y$ in X and $t_0, \ldots, t_{n-1}$ in $[T, \infty)$, denoted together by $(x_0, \ldots, x_n; t_0, \ldots, t_{n-1})$ such that,

$$d(\phi^{t_i}(x_i), x_{i+1}) < \varepsilon, \tag{16}$$

for $i = 0, 1, 2, \ldots, n - 1$.

Intuitively, an (ε, T) chain corresponds to the forward dynamics under flow ϕ connecting points $x, y \in X$, with slight perturbations allowed at each timestep (see Fig. 5 for an example). Note these deviations are allowed to occur at step-sizes T bounded away from 0, as otherwise the accumulation of perturbations could yield trajectories completely dissimilar to those induced by the original flow[54].

Definition 2.4.2 (Forward chain limit set). Let ϕ be a flow on a metric space (X, d). The forward chain limit set of $x \in X$ with respect to ϕ and d is the set,

$$\Omega^+(\phi, x) = \bigcap_{\varepsilon, T>0} \{y \in X | \ \exists \ \text{an}(\varepsilon, T)-\text{chain from } x \text{ to } y \text{ with respect to } \phi\}. \tag{17}$$

Definition 2.4.3 (Chain equivalent points). Let ϕ be a flow on a metric space (X, d). Two points $x, y \in X$ are chain equivalent with respect to ϕ and d if $y \in \Omega^+(\phi, x)$ and $x \in \Omega^+(\phi, y)$.

Definition 2.4.4 (Chain recurrent point). Let ϕ be a flow on a metric space (X, d). A point $x \in X$ is chain recurrent with respect to ϕ and d if x is chain equivalent to itself; i.e., there exists an (ε, T)-chain connecting x to itself for every $\varepsilon > 0$ and $T > 0$.

Chain recurrence can be understood as an orbit with slight perturbations allowed at each time step (see Fig. 5), which constitutes a new conception of "periodicity" with a very intuitive explanation in Computer Science terms: Imagine Alice is using a computer to simulate the trajectory of a dynamical system that induces a flow ϕ. Each iteration of the dynamical process computed by Alice, with a minimum step-size T, induces a rounding error ε. Consider an adversary, Bob, who can manipulate the result at each timestep within the ε-sphere of the actual result. If, regardless of ε or minimum step-size T, Bob can persuade Alice that her dynamical system starting from a point x returns back to this point in a finite number of steps, then this point is chain recurrent.

This new notion of "periodicity" (i.e., *chain recurrence*) leads to a corresponding notion of a "cycle" captured in the concept of *chain components*, defined below.

Definition 2.4.5 (Chain recurrent set). The chain recurrent set of flow ϕ, denoted $\mathcal{R}(\phi)$, is the set of all chain recurrent points of ϕ.

Definition 2.4.6 (Chain equivalence relation ~). Let the relation ~ on $\mathcal{R}(\phi)$ be defined by $x \sim y$ if and only if x is chain equivalent to y. This is an equivalence relation on the chain recurrent set $\mathcal{R}(\phi)$.

Definition 2.4.7 (Chain component). The equivalence classes in $\mathcal{R}(\phi)$ of the chain equivalence relation ~ are called the chain components of ϕ.

In the context of the Alice and Bob example, chain components are the maximal sets A such that for any two points $x, y \in A$, Bob can similarly persuade Alice that the flow ϕ induced by her dynamical system can get her from x to y in a finite number of steps. For example the matching pennies replicator dynamics (shown in Fig. 4a) have *one* chain component, consisting of the entire domain; in the context of the Alice and Bob example, the cyclical nature of the dynamics throughout the domain means that Bob can convince Alice that any two points may be connected using a series of finite perturbations ε, for all $\varepsilon > 0$ and $T > 0$. On the other hand, the coordination game replicator dynamics (shown in Fig. 4b) has five chain components corresponding to the fixed points (which are recurrent by definition): four in the corners, and one mixed strategy fixed point in the center. For a formal treatment of these examples, see[26,27].

Points in each chain component are transitive by definition. Naturally, the chain recurrent set $\mathcal{R}(\phi)$ can be partitioned into a (possibly infinitely many) number of chain components. In other words, chain components constitute the fundamental topological concept needed to define the irreducible behaviors we seek.

Conley's theorem. We now wish to characterize the role of chain components in the long-term dynamics of systems, such that we can evaluate the limiting behaviors of multi-agent interactions using our evolutionary

dynamical models. Conley's Fundamental Theorem of Dynamical Systems leverages the above perspective on "periodicity" (i.e., chain recurrence) and "cycles" (i.e., chain components) to decompose the domain of any dynamical system into two classes: 1) chain components, and 2) transient points.

We now need to formally define a complete Lyapunov function to introduce Conley's theorem. In game theoretic terms, one can understand this concept as the analog of a potential function, which strictly decreases along the dynamics trajectories in potential games, eventually leading to an equilibrium[55]. Correspondingly, under a complete Lyapunov function, the dynamics are led to chain recurrent sets (as opposed to equilibria). Formally:

Definition 2.4.8 (Complete Lyapunov function). Let ϕ be a flow on a metric space (X, d). A complete Lyapunov function for ϕ is a continuous function γ: $X \to \mathbb{R}$ such that,

1. $\gamma(\phi^t(x))$ is a strictly decreasing function of t for all $x \in X \backslash \mathcal{R}(\phi)$,
2. for all $x, y \in \mathcal{R}(\phi)$ the points x, y are in the same chain component if and only if $\gamma(x) = \gamma(y)$,
3. $\gamma(\mathcal{R}(\phi))$ is nowhere dense.

Conley's Theorem, the important result in topology that will form the basis of our solution concept and ranking scheme, is as follows:

Theorem 2.4.9 (Conley's Fundamental Theorem of Dynamical Systems[29], informal statement). The domain of any dynamical system can be decomposed into its (possibly infinitely many) chain components; the remaining points are transient, each led to the recurrent part by a Lyapunov function.

Conley's Theorem, critically, guarantees the existence of complete Lyapunov functions:

Theorem 2.4.10 Every flow on a compact metric space has a complete Lyapunov function[29].

In other words, the space X is decomposed into points that are chain recurrent and points that are led to the chain recurrent part in a gradient-like fashion with respect to a Lyapunov function that is guaranteed to exist. This implies that every game can be cast as a "potential" game if we consider chain recurrent sets as our solution concept.

Asymptotically stable sink chain components. Our objective is to investigate the likelihood of an agent being played in a K-wise meta-game by using a reasonable model of multi-agent evolution, such as the replicator dynamics. While chain components capture the limiting behaviors of dynamical systems (in particular, evolutionary dynamics that we seek to use for our multi-agent evaluations), they can be infinite in number (as mentioned in Section 2.4.1); it may not be feasible to compute or use them in practice within our evaluation scheme. To resolve this, we narrow our focus onto a particular class of chain components called *asymptotically stable sink chain components*, which we define in this section. Asymptotically stable sink chain components are a natural target for this investigation as they encode the possible "final" long term system; by contrast, we can escape out of other chain components via infinitesimally small perturbations. We prove in the subsequent section (Theorem 2.4.23, specifically) that, in the case of replicator dynamics and related variants, asymptotically stable sink chain components are finite in number; our desired solution concept is obtained as an artifact of this proof.

We proceed by first showing that the chain components of a dynamical system can be partially ordered by reachability through chains, and we focus on the *sinks* of this partial order. We start by defining a partial order on the set of chain components:

Definition 2.4.11 Let ϕ be a flow on a metric space and A_1, A_2 be chain components of the flow. Define the relation $A_1 \leq_C A_2$ to hold if and only if there exists $x \in A_2$ and $y \in A_1$ such that $y \in \Omega^+(\phi, x)$.

Intuitively, $A_1 \leq_C A_2$, if we can reach A_1 from A_2 with (ε, T)-chains for arbitrarily small ε and T.

Theorem 2.4.12 (Partial order on chain components). Let ϕ be a flow on a metric space and A_1, A_2 be chain components of the flow. Then the relation defined by $A_1 \leq_C A_2$ is a partial order.

Proof. Refer to the Supplementary Material for the proof. □

We will be focusing on minimal elements of this partial order, i.e., chain components A such that there does not exist any chain component B such that $B \leq_C A$. We call such chain components *sink chain components*.

Definition 2.4.13 (Sink chain components). A chain component A is called a sink chain component if there does not exist any chain component $B \neq A$ such that $B \leq_C A$.

We can now define the useful notion of asymptotically stable sink chain components, which relies on the notions of Lyapunov stable, asymptotically stable, and attracting sets.

Definition 2.4.14 (Lyapunov stable set). Let ϕ be a flow on a metric space (X, d). A set $A \subset X$ is Lyapunov stable if for every neighborhood O of A there exists a neighborhood O' of A such that every trajectory that starts in O' is contained in O; i.e., if $x \in O'$ then $\phi(t, x) \in O$ for all $t \geq 0$.

Definition 2.4.15 (Attracting set). Set A is attracting if there exists a neighborhood O of A such that every trajectory starting in O converges to A.

Definition 2.4.16 (Asymptotically stable set). A set is called asymptotically stable if it is both Lyapunov stable and attracting.

Definition 2.4.17 (Asymptotically stable sink chain component). Chain component A is called an asymptotically stable sink chain component if it is both a sink chain component and an asymptotically stable set.

Markov-Conley chains. Although we wish to study asymptotically stable sink chain components, it is difficult to do so theoretically as we do not have an exact characterization of their geometry and the behavior of the dynamics inside them. This is a rather difficult task to accomplish even experimentally. Replicator dynamics can be chaotic both in small and large games[52,56]. Even when their behavior is convergent for all initial conditions, the resulting equilibrium can be hard to predict and highly sensitive to initial conditions[57]. It is, therefore, not clear how to extract meaningful information even from many trial runs of the dynamics. These issues are exacerbated especially when games involve more than three or four strategies, where even visualization of trajectories becomes difficult.

Instead of studying the actual dynamics, a computationally amenable alternative is to use a discrete-time discrete-space approximation with similar limiting dynamics, but which can be directly and efficiently analyzed. We will start off by the most crude (but still meaningful) such approximations: a set of Markov chains whose state-space is the set of pure strategy profiles of the game. We refer to each of these Markov chains as a *Markov-Conley chain*, and prove in Theorem 2.4.23 that a finite number of them exist in any game under the replicator dynamics (or variants thereof).

Let us now formally define the Markov-Conley chains of a game, which rely on the notions of the response graph of a game and its sink strongly connected components.

Definition 2.4.18 (Strictly and weakly better response). Let $s_i, s_j \in \prod_k S^k$ be any two pure strategy profiles of the game, which differ in the strategy of a single player k. Strategy s_j is a strictly (respectively, weakly) better response than s_i for player k if her payoff at s_j is larger than (respectively, at least as large as) her payoff at s_i.

Definition 2.4.19 (Response graph of a game). The response graph of a game G is a directed graph whose vertex set coincides with the set of pure strategy profiles of the game, $\prod_k S^k$. Let $s_i, s_j \in \prod_k S^k$ be any two pure strategy profiles of the game. We include a directed edge from s_i to s_j if s_j is a weakly better response for player k as compared to s_i.

Definition 2.4.20 (Strongly connected components). The strongly connected components of a directed graph are the maximal subgraphs wherein there exists a path between each pair of vertices in the subgraph.

Definition 2.4.21 (Sink strongly connected components). The sink strongly connected components of a directed graph are the strongly connected components with no out-going edges.

The response graph of a game has a finite number of sink strongly connected components. If such a component is a singleton, it is a pure Nash equilibrium by definition.

Definition 2.4.22 (Markov-Conley chains (MCCs) of a game). A Markov-Conley chain of a game G is an irreducible Markov chain, the state space of which is a sink strongly connected component of the response graph associated with G. Many MCCs may exist for a given game G. In terms of the transition probabilities out of a node s_i of each MCC, a canonical way to define them is as follows: with some probability, the node self-transitions. The rest of the probability mass is split between all strictly and weakly improving responses of all players. Namely, the probability of strictly improving responses for all players are set equal to each other, and transitions between strategies of equal payoff happen with a smaller probability also equal to each other for all players.

When the context is clear, we sometimes overload notation and refer to the set of pure strategy profiles in a sink strongly connected component (as opposed to the Markov chain over them) as an MCC. The structure of the transition probabilities introduced in Definition 2.4.22 has the advantage that it renders the MCCs invariant under arbitrary positive affine transformations of the payoffs; i.e., the resulting theoretical and empirical insights are insensitive to such transformations, which is a useful desideratum for a game-theoretic solution concept. There may be alternative definitions of the transition probabilities that may warrant future exploration.

MCCs can be understood as a discrete approximation of the chain components of continuous-time dynamics (hence the connection to Conley's Theorem). The following theorem formalizes this relationship, and establishes finiteness of MCCs:

Theorem 2.4.23 Let ϕ be the replicator flow when applied to a K-player game. The number of asymptotically stable sink chain components is finite. Specifically, every asymptotically stable sink chain component contains at least one MCC; each MCC is contained in exactly one chain component.

Proof. Refer to the Supplementary Material for the proof. □

The notion of MCCs is thus used as a stepping stone, a computational handle that aims to mimic the long term behavior of replicator dynamics in general games. Similar results to Theorem 2.4.23 apply for several variants of replicator dynamics[13] as long as the dynamics are volume preserving in the interior of the state space, preserve the support of mixed strategies, and the dynamics act myopically in the presence of two strategies/options with fixed payoffs (i.e., if they have different payoffs then converge to the best, if they have the same payoffs then remain invariant).

From Markov-Conley chains to the discrete-time macro-model. The key idea behind the ordering of agents we wish to compute is that the evolutionary fitness/performance of a specific strategy should be reflected by how often it is being chosen by the system/evolution. We have established the solution concept of Markov-Conley chains (MCCs) as a discrete-time sparse-discrete-space analogue of the continuous-time replicator dynamics, which capture these long-term recurrent behaviors for general meta-games (see Theorem 2.4.23). MCCs are attractive from a computational standpoint: they can be found efficiently in all games by computing the

sink strongly connected components of the response graph, addressing one of the key criticisms of Nash equilibria. However, similar to Nash equilibria, even simple games may have many MCCs (e.g., five in the coordination game of Fig. 4b). The remaining challenge is, thus, to solve the MCC selection problem.

One of the simplest ways to resolve the MCC selection issue is to introduce noise in our system and study a stochastically perturbed version, such that the overall Markov chain is irreducible and, therefore, has a unique stationary distribution that can be used for our rankings. Specifically, we consider the following stochastically perturbed model: we choose a player k at random, and, if it is currently playing strategy s_i^k, we choose one of its strategies s_j^k at random and set the new system state to be $\varepsilon(s^k, s^{-k}) + (1 - \varepsilon)(s_j^k, s^{-k})$. Remarkably, these perturbed dynamics correspond closely to the macro-model introduced in Section 2.1.4 for a particularly large choice of ranking-intensity value α:

Theorem 2.5.1 In the limit of infinite ranking-intensity α, the Markov chain associated with the generalized multi-population model introduced in Section 2.1.4 coincides with the MCC.

Proof. Refer to the Supplementary Material for the proof. □

A low ranking-intensity ($\alpha \ll 1$) corresponds to the case of weak selection, where a weak mutant strategy can overtake a given population. A large ranking-intensity, on the other hand, ensures that the probability that a sub-optimal strategy overtakes a given population is close to zero, which corresponds closely to the MCC solution concept. In practice, setting the ranking-intensity to infinity may not be computationally feasible; in this case, the underlying Markov chain may be reducible and the existence of a unique stationary distribution (which we use for our rankings) may not be guaranteed. To resolve the MCC selection problem, we require a perturbed model, but one with a large enough ranking-intensity α such that it approximates an MCC, but small enough such that the MCCs remain connected. By introducing this perturbed version of Markov-Conley chains, the resulting Markov chain is now irreducible (per Theorem 2.1.2). The long-term behavior is thus captured by the unique stationary distribution under the large-α limit. Our so-called α-Rank evaluation method then corresponds to the ordering of the agents in this particular stationary distribution. The perturbations introduced here imply the need for a sweep over the ranking-intensity parameter α – a single hyperparameter – which we find to be computationally feasible across all of the large-scale games we analyze using α-Rank.

The combination of Theorem 2.4.23 and Theorem 2.5.1 yields a unifying perspective involving a chain of models of increasing complexity: the continuous-time replicator dynamics is on one end, our generalized discrete-time concept is on the other, and MCCs are the link in between (see Fig. 17).

Results

In the following we summarize our generalized ranking model and the main theoretical and empirical results. We start by outlining how the α-Rank procedure exactly works. Then we continue with illustrating α-Rank in a number of canonical examples. We continue with some deeper understanding of α-Rank's evolutionary dynamics model by introducing some further intuitions and theoretical results, and we end with an empirical validation of α-Rank in various domains.

α-Rank: evolutionary ranking of strategies. We first detail the α-Rank algorithm, then provide some insights and intuitions to further facilitate the understanding of our ranking method and solution concept.

Algorithm. Based on the dynamical concepts of chain recurrence and MCCs established, we now detail a descriptive method, titled α-Rank, for computing strategy rankings in a multi-agent interaction:

1. Construct the meta-game payoff table M^k for each population k from data of multi-agent interactions, or from running game simulations.
2. Compute the transition matrix C as outlined in Section 2.1.4. Per the discussions in Section 2.5, one must use a sufficiently large ranking-intensity value α in (4); this ensures that α-Rank preserves the ranking of strategies with closest correspondence to the MCC solution concept. Note that setting α arbitrarily high can result in numerical issues that make the representation of the Markov chain used in simulations reducible. As a large enough value is dependent on the domain under study, a useful heuristic is to conduct a sweep over α, starting from a small value and increasing it exponentially until convergence of rankings.
3. Compute the unique stationary distribution, π, of transition matrix C. Each element of the stationary distribution corresponds to the time the populations spend in a given strategy profile.
4. Compute the agent rankings, which correspond to the ordered masses of the stationary distribution π. The stationary distribution mass for each agent constitutes a 'score' for it (as might be shown, e.g., on a leaderboard).

α-Rank and MCCs as a solution concept: A paradigm shift. In this section, we elaborate on the differences between the Nash and MCC solution concepts. Our notion of a 'solution concept', informally, corresponds to a description of how agents will play a game. The MCC solution concept is not based on the idea of individual rationality, such as in Nash, but is rather biologically-conditioned, such as considered in evolutionary game theory[11,13]. As such, our solution concept of MCCs can be seen as a *descriptive* approach (in the sense of[58]) or *predictive* approach (in the sense of[10]), providing an understanding of the underlying dynamic behaviors as well as an understanding of what these behaviors converge to in the long-term. This is also where traditional game theory

differs from evolutionary game theory, in the sense that the former is normative and tells players how to play, while the latter is descriptive and relaxes some of the strong assumptions underpinning the Nash equilibrium concept.

We note that Nash has done double duty in game theory and remains a very important concept in multi-agent systems research. However, besides classical game theory making strong assumptions regarding the rationality of players involved in the interaction, there exist many fundamental limitations with the concept of a Nash equilibrium: intractability (computing a Nash is PPAD-complete), equilibrium selection, and the incompatibility of this static concept with the dynamic behaviors of agents in interacting systems. To compound these issues, even methods that aim to compute an approximate Nash are problematic: a typical approach is to use exploitability to measure deviation from Nash and as such use it as a method to closely approximate one; the problem with this is that it is also intractable for large games (typically the ones we are interested in), and there even still remain issues with using exploitability as a measure of strategy strength (e.g., see[59]). Overall, there seems little hope of deploying the Nash equilibrium as a solution concept for the evaluation of agents in general large-scale (empirical) games.

The concept of an MCC, by contrast, embraces the dynamical systems perspective, in a manner similar to evolutionary game theory. Rather than trying to capture the strategic behavior of players in an equilibrium, we deploy a dynamical system based on the evolutionary interactions of agents that captures and describes the long-term behavior of the players involved in the interaction. As such, our approach is descriptive rather than prescriptive, in the sense that it is not prescribing the strategies that one should play; rather, our approach provides useful information regarding the strategies that are evolutionarily non-transient (i.e., resistant to mutants), and highlights the remaining strategies that one might play in practice. To understand MCCs requires a shift away from the classical models described above for games and multi-agent interactions. Our new paradigm is to allow the dynamics to roll out and enable strong (i.e., non-transient) agents to emerge and weak (i.e., transient) agents to vanish naturally through their long-term interactions. The resulting solution concept not only permits an automatic ranking of agents' evolutionary strengths, but is powerful both in terms of computability and usability: our rankings are guaranteed to exist, can be computed tractably for any game, and involve no equilibrium selection issues as the evolutionary process converges to a unique stationary distribution. Nash tries to identify *static* single points in the simplex that capture simultaneous best response behaviors of agents, but comes with the range of complications mentioned above. On the other hand, the support of our stationary distribution captures the strongest non-transient agents, which may be interchangeably played by interacting populations and therefore constitute a *dynamic* output of our approach.

Given that both Nash and MCCs share a common foundation in the notion of a best response (i.e., simultaneous best responses for Nash, and the sink components of a best response graph for MCCs), it is interesting to consider the circumstances under which the two concepts coincide. There do, indeed, exist such exceptional circumstances: for example, for a potential game, every better response sequence converges to a (pure) Nash equilibrium, which coincides with an MCC. However, even in relatively simple games, differences between the two solution concepts are expected to occur in general due to the inherently dynamic nature of MCCs (as opposed to Nash). For example, in the Biased Rock-Paper-Scissors game detailed in Section 3.2.2, the Nash equilibrium and stationary distribution are not equivalent due to the cyclical nature of the game; each player's symmetric Nash is $\left(\frac{1}{16}, \frac{5}{8}, \frac{5}{16}\right)$, whereas the stationary distribution is $\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$. The key difference here is that whereas Nash is prescriptive and tells players which strategy mixture to use, namely $\left(\frac{1}{16}, \frac{5}{8}, \frac{5}{16}\right)$ assuming rational opponents, α-Rank is descriptive in the sense that it filters out evolutionary transient strategies and yields a ranking of the remaining strategies in terms of their long-term survival. In the Biased Rock-Paper-Scissors example, α-Rank reveals that all three strategies are equally likely to persist in the long-term as they are part of the same sink strongly connected component of the response graph. In other words, the stationary distribution mass (i.e., the α-Rank score) on a particular strategy is indicative of its resistance to being invaded by any other strategy, including those in the distribution support. In the case of the Biased Rock-Paper-Scissors game, this means that the three strategies are equally likely to be invaded by a mutant, in the sense that their outgoing fixation probabilities are equivalent. In contrast to our evolutionary ranking, Nash comes without any such stability properties (e.g., consider the interior mixed Nash in Fig. 4b). Even computing Evolutionary Stable Strategies (ESS)[13], a refinement of Nash equilibria, is intractable[60,61]. In larger games (e.g., AlphaZero in Section 3.4.2), the reduction in the number of agents that are resistant to mutations is more dramatic (in the sense of the stationary distribution support size being much smaller than the total number of agents) and less obvious (in the sense that more-resistant agents are not always the ones that have been trained for longer). In summary, the strategies chosen by our approach are those favored by evolutionary selection, as opposed to the Nash strategies, which are simultaneous best-responses.

Conceptual examples. We revisit the earlier conceptual examples of Rock-Paper-Scissors and Battle of the Sexes from Section 2.2 to illustrate the rankings provided by the α-Rank methodology. We use a population size of $m = 50$ in our evaluations.

Rock-Paper-Scissors. In the Rock-Paper-Scissors game, recall the cyclical nature of the discrete-time Markov chain (shown in Fig. 6a) for a fixed value of ranking-intensity parameter, α. We first investigate the impact of the ranking-intensity on overall strategy rankings, by plotting the stationary distribution as a function of α in Fig. 6b. The result is that the population spends $\frac{1}{3}$ of its time playing each strategy regardless of the value of α, which is in line with intuition due to the cyclical best-response structure of the game's payoffs. The Nash equilibrium, for comparison, is also $\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$. The α-Rank output Fig. 6c, which corresponds to a high value of α, thus indicates a tied ranking for all three strategies, also in line with intuition.

Biased Rock-Paper-Scissors. Consider now the game of Rock-Paper-Scissors, but with biased payoffs (shown in Fig. 7a). The introduction of the bias moves the Nash from the center of the simplex towards one of the corners, specifically $\left(\frac{1}{16}, \frac{5}{8}, \frac{5}{16}\right)$ in this case. It is worthwhile to investigate the corresponding variation of the stationary distribution masses as a function of the ranking-intensity α (Fig. 7c) in this case. As evident from the fixation probabilities (13) of the generalized discrete-time model, very small values of α cause the raw values of payoff to have a very low impact on the dynamics captured by discrete-time Markov chain; in this case, any mutant strategy has the same probability of taking over the population, regardless of the current strategy played by the population. This corresponds well to Fig. 7c, where small α values yield stationary distributions close to $\pi = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$.

As α increases, payoff values play a correspondingly more critical role in dictating the long-term population state; in Fig. 7c, the population tends to play Paper most often within this intermediate range of α. Most interesting to us, however, is the case where α increases to the point that our discrete-time model bears a close correspondence to the MCC solution concept (per Theorem 2.5.1). In this limit of large α, the striking outcome is that the stationary distribution once again converges to $\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$. Thus, α-Rank yields the high-level conclusion that in the long term, a monomorphic population playing any of the 3 given strategies can be completely and repeatedly displaced by a rare mutant, and as such assigns the same ranking to all strategies (Fig. 7d). This simple example illustrates perhaps the most important trait of the MCC solution concept and resulting α-Rank methodology: they capture the fundamental dynamical structure of games and long-term intransitivities that exist therein, with the rankings produced corresponding to the dynamical *strategy space consumption* or basins of attraction of strategies.

Battle of the sexes. We consider next an example of α-Rank applied to an asymmetric game – the Battle of the Sexes. Figure 8b plots the stationary distribution against ranking-intensity α, where we again observe a uniform stationary distribution corresponding to very low values of α. As α increases, we observe the emergence of two sink chain components corresponding to strategy profiles (O, O) and (M, M), which thus attain the top α-Rank scores in Fig. 8c. Note the distinct convergence behaviors of strategy profiles (O, M) and (M, O) in Fig. 8b, where the stationary distribution mass on the (M, O) converges to 0 faster than that of (O, M) for an increasing value of α. This is directly due to the structure of the underlying payoffs and the resulting differences in fixation probabilities. Namely, starting from profile (M, O), if either player deviates, that player increases their local payoff from 0 to 3. Likewise, if either player deviates starting from profile (O, M), that player's payoff increases from 0 to 2. Correspondingly, the fixation probabilities out of (M, O) are higher than those out of (O, M) (Fig. 8a), and thus the stationary distribution mass on (M, O) converges to 0 faster than that of (O, M) as α increases. We note that these low-α behaviors, while interesting, have no impact on the final rankings computed in the limit of large α (Fig. 8c). We refer the interested reader to[62] for a detailed analysis of the non-coordination components of the stationary distribution in mutualistic interactions, such as the Battle of the Sexes.

We conclude this discussion by noting that despite the asymmetric nature of the payoffs in this example, the computational techniques used by α-Rank to conduct the evaluation are essentially identical to the simpler (symmetric) Rock-Paper-Scissors game. This key advantage is especially evident in contrast to recent evaluation approaches that involve decomposition of a asymmetric game into multiple counterpart symmetric games, which must then be concurrently analyzed[7].

Theoretical properties of α-Rank. This section presents key properties related to the structure of the underlying discrete-time model used in α-Rank, and computational complexity of the ranking analysis.

Property 3.3.1 (Structure of C). Given strategy profile s_i corresponding to row i of C, the number of valid profiles it can transition to is $1 + \sum_k (|S^k| - 1)$ (i.e., either s_i self-transitions, or one of the populations k switches to a different monomorphic strategy). The sparsity of C is then,

$$1 - \frac{|S|(1 + \sum_k (|S^k| - 1))}{|S|^2}. \tag{18}$$

Therefore, for games involving many players and strategies, transition matrix C is large (in the sense that there exist $|S|$ states), but extremely sparse (in the sense that there exist only $1 + \sum_k (|S^k| - 1)$ outgoing edges from each state). For example, in a 6-wise interaction game where agents in each population have a choice over 4 strategies, C is 99.53% sparse.

Property 3.3.2 (Computational complexity of solving for π). The sparse structure of the Markov transition matrix C (as identified in Property 3.3.1) can be exploited to solve for the stationary distribution π efficiently; specifically, computing the stationary distribution can be formulated as an eigenvalue problem, which can be computed in cubic-time in the number of total pure strategy profiles. The α-Rank method is, therefore, tractable, in the sense that it runs in polynomial time with respect to the total number of pure strategies. This yields a major computational advantage, in stark contrast to conducting rankings by solving for Nash (which is PPAD-complete for general-sum games[21], which our meta-games may be).

Experimental validation. In this section we provide a series of experimental illustrations of α-Rank in a varied set of domains, including AlphaGo, AlphaZero Chess, MuJoCo Soccer, and both Kuhn and Leduc Poker. As evident in Table 1, the analysis conducted is extensive across multiple axes of complexity, as the domains considered include symmetric and asymmetric games with different numbers of populations and ranges of strategies.

AlphaGo. In this example we conduct an evolutionary ranking of AlphaGo agents based on the data reported in[1]. The meta-game considered here corresponds to a 2-player symmetric NFG with 7 AlphaGo agents: $AG(r)$, $AG(p)$, $AG(v)$, $AG(rv)$, $AG(rp)$, $AG(vp)$, and $AG(rvp)$, where r, v, and p respectively denote the combination of *rollouts*, *value networks*, and/or *policy networks* used by each variant. The corresponding payoffs are the win rates for each pair of agent match-ups, as reported in Table 9 of[1].

In Fig. 9c we summarize the rankings of these agents using the α-Rank method. α-Rank is quite conclusive in the sense that the top agent, $AG(rvp)$, attains all of the stationary distribution mass, dominating all other agents. Further insights into the pairwise agent interactions are revealed by visualizing the underlying Markov chain, shown in Fig. 9a. Here the population flows (corresponding to the graph edges) indicate which agents are more evolutionarily viable than others. For example, the edge indicating flow from $AG(r)$ to $AG(rv)$ indicates that the latter agent is stronger in the short-term of evolutionary interactions. Moreover, the stationary distribution (corresponding to high α values in Fig. 9b) reveals that all agents but $AG(rvp)$ are transient in terms of the long-term dynamics, as a monomorphic population starting from any other agent node eventually reaches $AG(rvp)$. In this sense, node $AG(rvp)$ constitutes an evolutionary stable strategy. We also see in Fig. 9a that no cyclic behaviors occur in these interactions. Finally, we remark that the recent work of[6] also conducted a meta-game analysis on these particular AlphaGo agents and drew similar conclusions to ours. The key limitation of their approach is that it can only directly analyze interactions between triplets of agents, as they rely on visualization of the continuous-time evolutionary dynamics on a 2-simplex. Thus, to draw conclusive results regarding the interactions of the full set of agents, they must concurrently conduct visual analysis of all possible 2-simplices (35 total in this case). This highlights a key benefit of α-Rank as it can succinctly summarize agent evaluations with minimal intermediate human-in-the-loop analysis.

AlphaZero. AlphaZero is a generalized algorithm that has been demonstrated to master the games of Go, Chess, and Shogi without reliance on human data[3]. Here we demonstrate the applicability of the α-Rank evaluation method to large-scale domains by considering the interactions of a large number of AlphaZero agents playing the game of chess. In AlphaZero, training commences by randomly initializing the parameters of a neural network used to play the game in conjunction with a general-purpose tree search algorithm. To synthesize the corresponding meta-game, we take a 'snapshot' of the network at various stages of training, each of which becomes an agent in our meta-game. For example, agent $AZ(27.5)$ corresponds to a snapshot taken at approximately 27.5% of the total number of training iterations, while $AZ(98.7)$ corresponds to one taken approximately at the conclusion of training. We take 56 of these snapshots in total. The meta-game considered here is then a symmetric 2-player NFG involving 56 agents, with payoffs again corresponding to the win-rates of every pair of agent match-ups. We note that there exist 27720 total simplex 2-faces in this dataset, substantially larger than those investigated in[6], which quantifiably justifies the computational feasibility of our evaluation scheme.

We first analyze the evolutionary strengths of agents over a sweep of ranking-intensity α (Fig. 10b). While the overall rankings are quite invariant to the value of α, we note again that a large value of α dictates the final α-Rank evaluations attained in Fig. 10c. To gain further insight into the inter-agent interactions, we consider the corresponding discrete-time evolutionary dynamics shown in Fig. 10a. Note that these interactions are evaluated using the entire 56-agent dataset, though visualized only for the top-ranked agents for readability. The majority of top-ranked agents indeed correspond to snapshots taken near the end of AlphaZero training (i.e., the strongest agents in terms of training time). Specifically, $AZ(99.4)$, which is the final snapshot in our dataset and thus the most-trained agent, attains the top rank with a score of 0.39, in contrast to the second-ranked $AZ(93.9)$ agent's score of 0.22. This analysis does reveal some interesting outcomes, however: agent $AZ(86.4)$ is not only ranked 5-th overall, but also higher than several agents with longer training time, including $AZ(88.8)$, $AZ(90.3)$, and $AZ(93.3)$.

We also investigate here the relationship between the α-Rank scores and Nash equilibria. A key point to recall is the equilibrium selection problem associated with Nash, as multiple equilibria can exist even in the case of two-player zero-sum meta-games. In the case of zero-sum meta-games, Balduzzi *et al.* show that there exists a unique *maximum entropy* (maxent) Nash equilibrium[63], which constitutes a natural choice that we also use in the below comparisons. For general games, unfortunately, this selection issue persists for Nash, whereas it does not for α-Rank due to the uniqueness of the associated ranking (see Theorem 2.1.2).

We compare the α-Rank scores and maxent Nash by plotting each throughout AlphaZero training in Fig. 11a,b, respectively; we also plot their difference in Fig. 11c. At a given training iteration, the corresponding horizontal slice in each plot visualizes the associated evaluation metric (i.e., α-Rank, maxent Nash, or difference of the two) computed for all agent snapshots up to that iteration. We first note that both evaluation methods reach a consensus that the strengths of AlphaZero agents generally increase with training, in the sense that only the latest agent snapshots (i.e., the ones closest to the diagonal) appear in the support of both α-Rank scores and Nash. An interesting observation is that less-trained agents sometimes reappear in the support of the distributions as training progresses; this behavior may even occur multiple times for a particular agent.

We consider also the quantitative similarity of α-Rank and Nash in this domain. Figure 11c illustrates that differences do exist in the sense that certain agents are ranked higher via one method compared to the other. More fundamentally, however, we note a relationship exists between α-Rank and Nash in the sense that they share a

common rooting in the concept of best-response: by definition, each player's strategy in a Nash equilibrium is a best response to the other players' strategies; in addition, α-Rank corresponds to the MCC solution concept, which itself is derived from the sink strongly-connected components of the game's response graph. Despite the similarities, α-Rank is a more refined solution concept than Nash in the sense that it is both rooted in dynamical systems and a best-response approach, which not only yields rankings, but also the associated dynamics graph (Fig. 10a) that gives insights into the long-term evolutionary strengths of agents. Beyond this, the critical advantage of α-Rank is its tractability for general-sum games (per Property 3.3.2), as well as lack of underlying equilibrium selection issues; in combination, these features yield a powerful empirical methodology with little room for user confusion or interpretability issues. This analysis reveals fundamental insights not only in terms of the benefits of using α-Rank to evaluate agents in a particular domain, but also an avenue of future work in terms of embedding the evaluation methodology into the training pipeline of agents involved in large and general games.

MuJoCo soccer. We consider here a dataset consisting of complex agent interactions in the continuous-action domain of MuJoCo soccer[5]. Specifically, this domain involves a multi-agent soccer physics-simulator environment with teams of 2 vs. 2 agents in the MuJoCo physics engine[64]. Each agent, specifically, uses a distinct variation of algorithmic and/or policy parameterization component (see[5] for agent specifications). The underlying meta-game is a 2-player NFG consisting of 10 agents, with payoffs corresponding to Fig. 2 of[5].

We consider again the variation of the stationary distribution as a function of ranking-intensity α (Fig. 12b). Under the large α limit, only 6 agent survive, with the remaining 4 agents considered transient in the long-term. Moreoever, the top 3 α-Ranked agents (*C*, *A*, and *B*, as shown in Fig. 12c) correspond to the strongest agents highlighted in[5], though α-Rank highlights 3 additional agents (*G*, *J*, and *F*) that are not in the top-rank set outlined in their work. An additional key benefit of our approach is that it can immediately highlight the presence of intransitive behaviors (cycles) in general games. Worthy of remark in this dataset is the presence of a large number of cycles, several of which are identified in Fig. 13. Not only can we identify these cycles visually, these intransitive behaviors are automatically taken into account in our rankings due to the fundamental role that recurrence plays in our underlying solution concept. This is in contrast to the Elo rating (which is incapable of dealing with intransitivities), the replicator dynamics (which are limited in terms of visualizing such intransitive behaviors for large games), and Nash (which is a static solution concept that does not capture dynamic behavior).

Kuhn poker. We next consider games wherein the inherent complexity is due to the number of players involved. Specifically, we consider Kuhn poker with 3 and 4 players, extending beyond the reach of prior meta-game evaluation approaches that are limited to pairwise asymmetric interactions[6]. Kuhn poker is a small poker game in which every player starts with 2 chips, antes 1 chip to play, and receives a single card face down from a deck of size $n+1$ (one card remains hidden). Then the players can bet (raise/call) by adding their remaining chip to the pot, or can pass (check/fold) until all players are either in (contributed to the pot) or out (folded, passed after a raise). The player with the highest-ranked card that has not folded wins the pot. The two-player game is known to have a continuum of strategies, which could have fairly high support, that depends on a single parameter: the probability that the first player raises with the highest card[65]. The three-player game has a significantly more complex landscape[66]. The specific rules used for the three and four player variants can be found in[67], [Section 4.1].

Here, our meta-game dataset consists of several (fixed) rounds of extensive-form fictitious play (specifically, XFP from[68]): in round 0, the payoff corresponding to strategy profile (0, 0, 0) in each meta-game of 3-player Kuhn corresponds to the estimated payoff of each player using uniform random strategies; in fictitious play round 1, the payoff entry (1, 1, 1) corresponds to each player playing an approximate best response to the other players' uniform strategies; in fictitious play round 2, entry (2, 2, 2) corresponds to each playing an approximate best response to the other players' uniform mixtures over their round 0 strategies (uniform random) and round 1 oracle strategy (best response to random); and so on. Note, especially, that oracles at round 0 are likely to be dominated (as they are uniform random). In our dataset, we consider two asymmetric meta-games, each involving 3 rounds of fictitious play with 3 and 4 players (Figs 14 and 15, respectively).

Of particular note are the total number of strategy profiles involved in these meta-games, 64 and 256 respectively for the 3 and 4 player games – the highest considered in any of our datasets. Conducting the evaluation using the replicator-dynamics based analysis of[6] can be quite tedious as all possible 2-face simplices must be considered for *each* player. Instead, here the α-Rankings follow the same methodology used for all other domains, and are summarized succinctly in Figs 14c and 15c. In both meta-games, the 3-round fictitious play strategies ((3, 3, 3) and (3, 3, 3, 3), respectively) are ranked amongst the top-5 strategies.

Leduc poker. The meta-game we consider next involves agents generated using the Policy Space Response Oracles (PSRO) algorithm[30]. Specifically, PSRO can be viewed as a generalization of fictitious play, which computes approximate responses ("oracles") using deep reinforcement learning, along with arbitrary meta-strategy solvers; here, PSRO is applied to the game of Leduc poker. Leduc poker involves a deck of 6 cards (jack, queen, and king in two suits). Players have a limitless number of chips. Each player antes 1 chip to play and receives an initial private card; in the first round players can bet a fixed amount of 2 chips, in the second round can bet 4 chips, with a maximum of two raises in each round. Before the second round starts, a public card is revealed. The corresponding meta-game involves 2 players with 3 strategies each, which correspond to the first three epochs of the PSRO algorithm. Leduc poker is a commonly used benchmark in the computer poker literature[69]: our implementation contains 936 information states (approximately 50 times larger then 2-player Kuhn poker), and is non-zero sum due to penalties imposed by selecting of illegal moves, see [30, Appendix D.1] for details.

We consider in Fig. 16a the Markov chain corresponding to the PSRO dataset, with the corresponding α-Rank yielding profile $(0, 0)$ as the top-ranked strategy, which receives 1.0 of the stationary distribution mass and essentially consumes the entire strategy space in the long-term of the evolutionary dynamics. This corresponds well to the result of[6], which also concluded that this strategy profile consumes the entire strategy space under the replicator dynamics; in their approach, however, an equilibrium selection problem had to be dealt with using human-in-the-loop intervention due to the population-wise dynamics decomposition their approach relies on. Here, we need no such intervention as α-Rank directly yields the overall ranking of all strategy profiles.

Discussion

We introduced a general descriptive multi-agent evaluation method, called α-*Rank*, which is practical and general in the sense that it is easily applicable in complex game-theoretic settings, including K-player asymmetric games that existing meta-game evaluation methods such as[6,7] cannot feasibly be applied to. The techniques underlying α-Rank were motivated due to the fundamental incompatibility identified between the dynamical processes typically used to model interactions of agents in meta-games, and the Nash solution concept typically used to draw conclusions about these interactions. Using the Nash equilibrium as a solution concept for meta-game evaluation in these dynamical models is in many ways problematic: computing a Nash equilibrium is not only computationally difficult[21,22], and there are also intractable equilibrium selection issues even if Nash equilibria can be computed[23–25]. α-Rank, instead, is theoretically-grounded in a novel solution concept called Markov-Conley chains (MCCs), which are inherently dynamical in nature. A key feature of α-Rank is that it relies on only a single hyper-parameter, its ranking-intensity value α, with sufficiently high values of α (as determined via a parameter sweep) yielding closest correspondence to MCCs.

The combination of MCCs and α-Rank yields a principled methodology with a strong evolutionary interpretation of agent rankings, as outlined in Fig. 17; this overarching perspective considers a spectrum of evolutionary models of increasing complexity. On one end of the spectrum, the continuous-time dynamics micro-model provides detailed insights into the simplex, illustrating flows, attractors, and equilibria of agent interactions. On the other end, the discrete-time dynamics macro-model provides high-level insights of the time limit behavior of the system as modeled by a Markov chain over interacting agents. The unifying link between these models is the MCC solution concept, which builds on the dynamical theory foundations of Conley[29] and the topological concept of chain components. We provided both scalability properties and theoretical guarantees for our ranking method. Finally, we evaluated the approach on an extensive range of meta-game domains including AlphaGo[1], AlphaZero[3], MuJoCo Soccer[5], and Poker[30], which exhibit a range of complexities in terms of payoff asymmetries, number of players, and number of agents involved. We strongly believe that the generality of α-Rank will enable it to play an important role in evaluation of AI agents, e.g., on leaderboards. More critically, we believe that the computational feasibility of the approach, even when many agents are involved (e.g., AlphaZero), makes its integration into the agent *training* pipeline a natural avenue for future work.

References

1. Silver, D. *et al*. Mastering the game of Go with deep neural networks and tree search. *Nature* **5290**(7587), 484–489 (2016).
2. Silver, D. *et al*. Mastering the game of Go without human knowledge. *Nature* **550**, 354–359 (2017).
3. Silver, D. *et al*. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science* **3620**(6419), 1140–1144 (2018).
4. Moravčík, M. *et al*. DeepStack: Expert-level artificial intelligence in heads-up no-limit poker. *Science* 356(6337):508–513, ISSN 0036-8075 (2017).
5. Liu, S. *et al*. Emergent coordination through competition. In *International Conference on Learning Representations*, https://openreview.net/forum?id=BkG8sjR5Km (2019).
6. Tuyls, K., Perolat, J., Lanctot, M., Leibo, J. Z. & Graepel, T. A Generalised Method for Empirical Game Theoretic Analysis. In *AAMAS*, *Stockholm*, *Sweden* (2018).
7. Tuyls, K. *et al*. Symmetric decomposition of asymmetric games. *Scientific Reports* **80**(1), 1015 (2018).
8. Walsh, W. E., Das, R., Tesauro, G. & Kephart, J. O. Analyzing complex strategic interactions in multi-agent games. In *AAAI-02 Workshop on Game Theoretic and Decision Theoretic Agents*, *2002* (2002).
9. Wellman, M. P. Methods for empirical game-theoretic analysis. In Proceedings, The Twenty-First National Conference on Artificial Intelligence and the Eighteenth Innovative Applications of Artificial Intelligence Conference, July 16–20, 2006, Boston, Massachusetts, USA, pages 1552–1556 (2006).
10. Tuyls, K. & Parsons, S. What evolutionary game theory tells us about multiagent learning. *Artif. Intell.* **1710**(7), 406–416 (2007).
11. Gintis, H. Game theory evolving (2nd edition). (*University Press*, *Princeton NJ*, 2009).
12. Hofbauer, J. Evolutionary dynamics for bimatrix games: A Hamiltonian system? *J. of Math. Biology* **34**, 675–688 (1996).
13. Weibull, J. *Evolutionary game theory* (MIT press, 1997).
14. Zeeman, E. C. Population dynamics from game theory. *Lecture Notes in Mathematics*, *Global theory of dynamical systems* **819** (1980).
15. Zeeman, E. C. Dynamics of the evolution of animal conflicts. *Theoretical Biology* **89**, 249–270 (1981).
16. Santos, F. C., Pacheco, J. M. & Skyrms, B. Co-evolution of pre-play signaling and cooperation. *Journal of Theoretical Biology* **2740**(1), 30–35 (2011).
17. Segbroeck, S. V., Pacheco, J. M., Lenaerts, T. & Santos, F. C. Emergence of fairness in repeated group interactions. *Physical Review Letters* **108**, 158104 (2012).
18. Traulsen, A., Claussen, J. C. & Hauert, C. Coevolutionary dynamics: from finite to infinite populations. *Physical review letters* **950**(23), 238701 (2005).
19. Traulsen, A., Nowak, M. A. & Pacheco, J. M. Stochastic dynamics of invasion and fixation. *Phys. Rev. E* **74**, 011909 (2006).
20. Veller, C. & Hayward, L. K. Finite-population evolution with rare mutations in asymmetric games. *Journal of Economic Theory* **162**, 93–113 (2016).
21. Daskalakis, C., Goldberg, P. W. & Papadimitriou, C. H. The complexity of computing a Nash equilibrium. In *Proceedings of the 38th Annual ACM Symposium on Theory of Computing*, *Seattle*, *WA*, *USA*, *May 21–23*, *2006*, pages 71–78 (ACM Press, 2006).
22. von Stengel, B. Computing equilibria for two-person games. In *Handbook of Game Theory with Economic Applications*, volume 3, pages 1723–1759 (Elsevier, 2002).
23. Avis, D., Rosenberg, G., Savani, R. & von Stengel, B. Enumeration of nash equilibria for two-player games. *Economic Theory* **42**, 9–37 (2010).

24. Goldberg, P. W., Papadimitriou, C. H. & Savani, R. The complexity of the homotopy method, equilibrium selection, and Lemke-Howson solutions. *ACM Transactions on Economics and Computation* **10**(2), 9 (2013).
25. Harsanyi, J. & Selten, R. *A General Theory of Equilibrium Selection in Games*, volume 1 (The MIT Press, 1 edition, 1988).
26. Papadimitriou, C. & Piliouras, G. From Nash equilibria to chain recurrent sets: Solution concepts and topology. In *Proceedings of the 2016* ACM *Conference on Innovations in Theoretical Computer Science*, ITCS '16, pages 227–235, New York, NY, USA (ACM, ISBN 978-1-4503-4057-1 (2016).
27. Papadimitriou, C. & Piliouras, G. Game dynamics as the meaning a game. *Sigecom Exchanges* **16**, 2 (2018).
28. Kakutani, S. A generalization of Brouwer's fixed point theorem. *Duke Mathematical Journal* **80**(3), 457–459 (1941).
29. Conley, C. C. *Isolated invariant sets and the Morse index*. Number 38 (American Mathematical Soc., 1978).
30. Lanctot, M. *et al*. A unified game-theoretic approach to multiagent reinforcement learning. In *Advances in Neural Information Processing Systems 30*, pages 4190–4203 (2017).
31. Cressman, R. *Evolutionary Dynamics and Extensive Form Games*. (The MIT Press, 2003).
32. Hofbauer, J. J. & Sigmund, K. *Evolutionary games and population dynamics*. (Cambridge University Press, 1998).
33. Evans, R. C. & Harris, F. H. De. B. A bayesian analysis of free rider metagames. *Southern Economic Journal* **490**(1), 137–149 (1982).
34. Schuster, P. & Sigmund, K. Replicator dynamics. *Journal of Theoretical Biology* 1000 (3): 533–538, ISSN 0022-5193, https://doi.org/10.1016/0022-5193(83)90445-9, http://www.sciencedirect.com/science/article/pii/0022519383904459 (1983).
35. Taylor, P. & Jonker, L. Evolutionarily stable strategies and game dynamics. *Mathematical Biosciences* **40**, 145–156 (1978).
36. Bloembergen, Daan, Tuyls, Karl, Hennes, Daniel & Kaisers, Michael Evolutionary dynamics of multi-agent learning: A survey. *J. Artif. Intell. Res. (JAIR)* **53**, 659–697 (2015).
37. Fudenberg, D. & Imhof, L. A. Imitation processes with small mutations. *Journal of Economic Theory* **1310**(1), 251–262 (2006).
38. Nowak, M. A. & Sigmund, K. Evolutionary dynamics of biological games. *Science* **3030**(5659), 793–799 (2004).
39. Traulsen, A., Pacheco, J. M. & Imhof, L. A. Stochasticity and evolutionary stability. *Phys. Rev. E* **74**, 021905 (2006).
40. Claussen, J. C. Discrete stochastic processes, replicator and Fokker-Planck equations of coevolutionary dynamics in finite and infinite populations. *arXiv preprint arXiv:0803.2443* (2008).
41. Taylor, H. M. & Karlin, S. *An Introduction To Stochastic Modeling* (Academic Press, third edition edition, 1998).
42. Daskalakis, C., Frongillo, R., Papadimitriou, C., Pierrakos, G. & Valiant, G. On learning algorithms for Nash equilibria. *Algorithmic Game Theory*, pages 114–125 (2010).
43. Hart, S. & Mas-Colell, A. Uncoupled dynamics do not lead to nash equilibrium. *American Economic Review* **930**(5), 1830–1836 (2003).
44. Viossat, Y. The replicator dynamics does not lead to correlated equilibria. *Games and Economic Behavior* **590**(2), 397–407 (2007).
45. Piliouras, G. & Schulman, L. J. Learning dynamics and the co-evolution of competing sexual species. *arXiv preprint arXiv:1711.06879* (2017).
46. Sandholm, W. H. *Population Games and Evolutionary Dynamics*. Economic Learning and Social Evolution, ISBN 9780262288613 (MIT Press, 2010).
47. Gaunersdorfer, A. & Hofbauer, J. Fictitious play, shapley polygons, and the replicator equation. *Games and Economic Behavior* **11**, 279–303 (1995).
48. Kleinberg, R., Ligett, K., Piliouras, G. & Tardos, É. Beyond the Nash equilibrium barrier. In *Symposium on Innovations in Computer Science (ICS)* (2011).
49. Palaiopanos, G., Panageas, I. & Piliouras, G. Multiplicative weights update with constant step-size in congestion games: Convergence, limit cycles and chaos. In *NIPS* (2017).
50. Sandholm, W. H. *Population games and evolutionary dynamics*. (MIT press, 2010).
51. Wagner, E. The explanatory relevance of nash equilibrium: One-dimensional chaos in boundedly rational learning. *Philosophy of Science* **800**(5), 783–795 (2013).
52. Sato, Y., Akiyama, E. & Farmer, J. D. Chaos in learning a simple two-person game. *Proceedings of the National Academy of Sciences* **990**(7), 4748–4751 (2002).
53. Alongi, J. M. & Nelson, G. S. *Recurrence and Topology*, volume 85 (American Mathematical Soc., 2007).
54. Norton, D. E. The fundamental theorem of dynamical systems. *Commentationes Mathematicae Universitatis Carolinae* **360**(3), 585–597 (1995).
55. Monderer, D. & Shapley, L. S. Potential Games. *Games and Economic Behavior* **14**, 124–143 (1996).
56. Galla, T. & Farmer, J. D. Complex dynamics in learning complicated games. *Proceedings of the National Academy of Sciences* **1100**(4), 1232–1236 (2013).
57. Panageas, I. & Piliouras, G. Average case performance of replicator dynamics in potential games via computing regions of attraction. In *Proceedings of the 2016 ACM Conference on Economics and Computation*, pages 703–720 (ACM, 2016).
58. Shoham, Y., Powers, R. & Grenager, T. If multi-agent learning is the answer, what is the question? *Artificial Intelligence* **1710**(7), 365–377 (2007).
59. Davis, T., Burch, N. & Bowling, M. Using response functions to measure strategy strength. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, *July 27–31*, *2014*, *Québec City*, *Québec*, *Canada*., pages 630–636 (2014).
60. Conitzer, V. The exact computational complexity of evolutionarily stable strategies. *CoRR*, abs/1805.02226 (2018).
61. Etessami, K. & Lochbihler, A. The computational complexity of evolutionarily stable strategies. *International Journal of Game Theory* (2008).
62. Veller, C., Hayward, L. K., Hilbe, C. & Nowak, M. A. The red queen and king in finite populations. *Proceedings of the National Academy of Sciences* **1140**(27), E5396–E5405 (2017).
63. Balduzzi, D., Tuyls, K., Perolat, J. & Graepel, T. Re-evaluating Evaluation. *arXiv*, 0 (1806.02643) (2018).
64. Todorov, E., Erez, T. & Tassa, Y. Mujoco: A physics engine for model-based control. In *IROS* (2012).
65. Southey, F., Hoehn, B. & Holte, R. C. Effective short-term opponent exploitation in simplified poker. *Machine Learning* **740**(2), 159–189 (2009).
66. Szafron, D., Gibson, R. & Sturtevant, N. A parameterized family of equilibrium profiles for three-player Kuhn poker. In *Proceedings of the Twelfth International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 247–254 (2013).
67. Lanctot, M. Further developments of extensive-form replicator dynamics using the sequence-form representation. In *Proceedings of the Thirteenth International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, pages 1257–1264 (2014).
68. Heinrich, J., Lanctot, M. & Silver, D. Fictitious self-play in extensive-form games. In *Proceedings of the 32nd International Conference on Machine Learning (ICML 2015)* (2015).
69. Southey, F. *et al*. Bayes' bluff: Opponent modelling in poker. In *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence (UAI 2005)* (2005).

Acknowledgements

We are very grateful to G. Ostrovski, T. Graepel, E. Hughes, Y. Bachrach, K. Kavukcuoglu, D. Silver, T. Hubert, J. Schrittwieser, S. Liu, G. Lever, and D. Bloembergen for helpful comments, discussions, and for making available datasets used in this document. Christos Papadimitriou acknowledges NSF grant 1408635 "Algorithmic Explorations of Networks, Markets, Evolution, and the Brain", and NSF grant 1763970 to Columbia University. Georgios Piliouras acknowledges SUTD grant SRG ESD 2015 097, MOE AcRF Tier 2 Grant 2016-T2-1-170, grant

PIE-SGP-AI-2018-01 and NRF 2018 Fellowship NRF-NRFF2018-07. In memory of Catherine Renoir, a close friend and colleague at DeepMind.

Author Contributions

S.O., C.P., G.P. and K.T. designed the research, theoretical contributions, and implemented the experiments. S.O., C.P., G.P., K.T., M.R., J.-B.L., W.C., M.L., J.P. and R.M. analyzed the results and wrote and reviewed the paper.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-019-45619-9.

Competing Interests: The authors declare no competing interests.

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