SOCIAL SCIENCES

Best reply structure and equilibrium convergence in generic games

Marco Pangallo[1,2]*, Torsten Heinrich[1,2,3], J. Doyne Farmer[1,2,4,5]

Game theory is widely used to model interacting biological and social systems. In some situations, players may converge to an equilibrium, e.g., a Nash equilibrium, but in other situations their strategic dynamics oscillate endogenously. If the system is not designed to encourage convergence, which of these two behaviors can we expect a priori? To address this question, we follow an approach that is popular in theoretical ecology to study the stability of ecosystems: We generate payoff matrices at random, subject to constraints that may represent properties of real-world games. We show that best reply cycles, basic topological structures in games, predict nonconvergence of six well-known learning algorithms that are used in biology or have support from experiments with human players. Best reply cycles are dominant in complicated and competitive games, indicating that in this case equilibrium is typically an unrealistic assumption, and one must explicitly model the dynamics of learning.

INTRODUCTION

Game theory is a set of mathematical models for strategic interactions between decision makers (*1*) with applications to diverse problems such as the emergence of cooperation (*2*), language formation (*3*), and congestion on roads and on the internet (*4*). The same mathematical tools are used to model evolution in ecology and population biology (*5*). A long-standing question is whether players will converge to an equilibrium as they learn by playing a game repeatedly. Little is known about the general case, as the answer depends on the properties of the game and on the learning algorithm. Here, we introduce a formalism that we call best reply structure that gives a rough measure of the convergence probability in any two-player normal-form game for a wide class of learning algorithms. Analogies that illustrate the usefulness of our approach in other fields are the theory of qualitative stability in ecology (*6*) and the use of the Reynolds number to understand turbulence in fluid dynamics (*7*).

The standard approach to the problem of equilibrium convergence in game theory is to focus on classes of games with special mathematical properties, selected as stylized models of real-world scenarios. For example, in potential games (*8*), all differences in payoffs for unilaterally changing strategy can be expressed using a global potential function; congestion games (*4*) belong to this class, so potential games can be used as a stylized model of traffic. Most learning algorithms converge to a Nash or correlated equilibrium in potential games, as well as in dominance-solvable (*9*), coordination (*10*), supermodular (*11*), and "weakly acyclic" (*12*) games. Studying classes of games with special properties is really useful in situations such as mechanism design, in which one can choose the game to be played. In these situations, it may also be possible to design the learning algorithm that the players are going to use, and choose the one that most likely will converge to equilibrium. Examples include online auctions and internet routing (*13*).

There are other problems, however, where the game and the learning algorithm are not designed, but rather are dictated by the intrinsic nature of the setup. For example, a financial market can be viewed as a game with many possible actions corresponding to the assets that traders can buy or sell. The outcome might converge to an equilibrium, or it might endogenously fluctuate. If the system is not designed to encourage convergence, which of these two behaviors should we expect?

To address this issue, we follow an approach that has been extremely productive in theoretical ecology and developmental biology and is widespread in physics. An example of this approach is the seminal paper in theoretical ecology by May (*14*). He studied an ensemble of randomly generated predator-prey interactions, which he took as a null model of a generic ecosystem. His key result was that random ecosystems tend to become more unstable as they grow larger. Of course, as May was well aware, real ecosystems are not random; rather, they are shaped by evolutionary selection and other forces. Many large ecosystems have existed for long periods of time, suggesting that they are in fact stable. Thus, this contradiction indicated that real ecosystems are not typical members of the random ensemble used in May's model, raising the important question of precisely how these ecosystems are atypical, and how and why they evolved to become stable. Forty-five years later, properly answering this question remains a subject of active research. For example, Johnson *et al.* (*15*) recently found that real ecosystems have a property that they call trophic coherence, and showed that incorporating this property as a constraint on the ensemble of randomly generated ecosystems ensures stability.

Here, we apply a similar approach to game theory, taking an ensemble of randomly generated two-player games as a null model. For reasons of tractability, we study normal-form games, taking advantage of the fact that it is possible to systematically enumerate all possible games. The null model is refined by adding constraints that can be varied to understand their effect on convergence to equilibrium. Here, we study in detail a parameter Γ that tunes the correlation of the payoffs to the two players. This regulates the intensity of competition in a game and encompasses zero-sum games as a special case for $\Gamma = -1$. We also sketch how one might construct other constraints, for example, to study deviations from potential games. With this approach, it is possible to see how deviations from particular classes of games affect the stability of the learning dynamics.

Randomly generated games and general learning algorithms do not have mathematical properties that allow exact solutions. To overcome this limitation, we develop a formalism to obtain the approximate probability of convergence as a function of a simple indicator. An analogy to

[1]Institute for New Economic Thinking at the Oxford Martin School, University of Oxford, Oxford OX2 6ED, UK. [2]Mathematical Institute, University of Oxford, Oxford OX1 3LP, UK. [3]Department for Business Studies and Economics, University of Bremen, 28359 Bremen, Germany. [4]Computer Science Department, University of Oxford, Oxford OX1 3QD, UK. [5]Santa Fe Institute, Santa Fe, NM 87501, USA.
*Corresponding author. Email: marco.pangallo@maths.ox.ac.uk

fluid dynamics clarifies how these approximate solutions can be useful. As a fluid is driven away from equilibrium, it typically makes a transition from stable (or laminar) flow to unstable (or turbulent) flow. There exist no analytical solutions for the Navier-Stokes equations that describe fluid dynamics, but this transition can nonetheless be crudely characterized in terms of a nondimensional parameter called the Reynolds number (*7*). A larger Reynolds number means a higher likelihood of turbulence. Although this prediction is imprecise—it is just a rule of thumb—it is nonetheless very useful. Our analogous estimate for games does not have a simple closed form, but it has similar predictive power. Another analogy to our approach is the theory of qualitative stability in theoretical ecology (*6*). Many models in ecology consider the magnitude of the interactions between different species in a food web. For example, these models consider how much grass is eaten by rabbits and how many rabbits are eaten by foxes. The qualitative stability approach instead considers only the sign of the predator-prey relations—that rabbits eat grass and foxes eat rabbits. This makes it possible to obtain an approximate assessment of the stability of an ecosystem just from the topological properties of its food web. As with the theory of qualitative stability, our approach only depends on the topological properties of the game and not on the details of the payoffs.

Our formalism is based on best reply dynamics, under which each player myopically responds with the best reply to her opponent's last action. Best reply dynamics is well known in game theory and is widely used to develop intuition about learning, but we use it in a new way, to obtain the approximate probability of convergence in generic games. Under best reply dynamics, the system will asymptotically either converge to a fixed point, corresponding to a pure strategy Nash equilibrium, or be trapped on a cycle. We consider a very simple indicator of nonconvergence in a game based on the relative size of the best reply cycles versus the fixed points of a game. Note that we are not assuming that players follow best reply dynamics. Rather, we hypothesize that the best reply structure of the payoff matrix constitutes a first-order skeleton, forming the backbone of the game the players are trying to learn, which is useful for understanding the convergence of many learning algorithms.

To test this hypothesis, we choose a set of learning algorithms that have support from experiments with human players. These are reinforcement learning, fictitious play, experience-weighted attraction with and without noise, and level-*k* learning. We also include (two-population) replicator dynamics for its importance in ecology and population biology. Our measure based on best reply dynamics predicts the nonconvergence frequency of each of these algorithms with $R^2 \geq 0.78$.

Here, we want to stress that our goal is descriptive rather than normative. In mechanism design, people or machines use algorithms that are designed to have good convergence properties. For example, some algorithms converge to correlated equilibria, a generalization of Nash equilibrium that allows players to coordinate on a common signal in all games (*16*–*20*). One such algorithm is regret matching (*17*), in which players consider all past history of play and calculate what their payoff would have been had they taken any other action. While these algorithms might be feasibly executed by a machine or by a human with sufficient record-keeping ability, it seems unlikely that they would actually be used by real people unless they were specifically trained to do so. To the best of our knowledge, these algorithms only have indirect empirical support. We focus on algorithms that have been tested experimentally. When they reach a fixed point, these algorithms converge to a Nash equilibrium , or a point near one, rather than a more general correlated equilibrium.

After showing that the best reply structure formalism works, we analyze how best reply cycles or fixed points vary as two properties of the game change. We define how complicated a game is based on the number of actions N available to the players. A simple game has only a few actions, and a complicated game has many actions. The competitiveness of the game is defined by the correlation Γ between the payoffs of the two players. The more negative the correlation, the more competitive the game. The relative share of best reply cycles versus fixed points tracks the convergence frequency of the six algorithms we consider as we vary these two properties of the game, with the exception of fictitious play in competitive games.

We show that at one end of the spectrum, games that are simple and noncompetitive are unlikely to have cycles, while at the other end, games that are complicated and competitive are likely to have cycles. The classes of games that we mentioned before, i.e., potential, dominance-solvable, coordination, supermodular, and weakly acyclic games, are acyclic by construction (*8*–*12*). Any of these classes might be typical members of the ensemble of simple noncompetitive games, where acyclic behavior is common, but they are certainly not typical for games that are complicated and competitive. These results match the intuition that complicated games are harder to learn and it is harder for players to coordinate on an equilibrium when one player's gain is the other player's loss. Our formalism makes this quantitative. For example, with two actions per player and no correlation between the payoffs to the two players ($\Gamma = 0$), acyclic games are about 85% of the total. However, with $N = 10$ and $\Gamma = -0.7$, they make up only 2.7% of the total.

We also show how it is possible to use the best reply formalism to study the stability of a given class of games, such as potential games, to understand their stability under deviations from the given class. We show that the behavior can be nonlinear, e.g., small perturbations of potential games cause a disproportionately large increase in nonconvergence.

In the case of uncorrelated payoffs, $\Gamma = 0$, we use combinatorial methods inspired by statistical mechanics to analytically compute the frequency of best reply cycles of different lengths. The idea of using methods inspired from statistical mechanics is not new in game theory (*21*). Previous research has quantified properties of pure strategy Nash equilibrium (*22*–*25*), mixed strategy equilibria (*26*, *27*), and Pareto equilibria (*28*), but we are the first to quantify the frequency and length of best reply cycles and show their relevance for learning. The frequency of convergence for experience-weighted attraction in random games was previously studied by Galla and Farmer (*29*) in the limit as $N \to \infty$ using different methods; here, we extend this to arbitrary N, study a diverse set of learning algorithms, and provide deeper insight into the origins of instability. The formalism we introduce can be extended in many directions and used in different fields. For example, our results are also related to the stability of food webs (*6*, *14*) through replicator dynamics and can be mapped into Boolean networks (*30*).

When convergence to equilibrium fails, we often observe chaotic learning dynamics (*29*, *31*). When this happens, the players do not converge to any sort of intertemporal "chaotic equilibrium" (*32*–*34*) in the sense that their expectations do not match the outcomes of the game even in a statistical sense. In many cases, the resulting attractor is high dimensional, making it difficult for a "rational" player to outperform other players by forecasting their actions using statistical methods. Once at least one player systematically deviates from equilibrium, learning and heuristics can outperform equilibrium thinking (*35*) and can be a better description for the behavior of players.

We begin by developing our best reply framework. We next show how it can be used to predict the frequency of nonconvergence of the six learning algorithms that we study here, first presenting some arguments giving some intuition about why this works and then providing more

quantitative evidence. We then study whether best reply cycles become prevalent as some properties of the games change, illustrating the effect of two different constraints that represent deviations from well-known classes of games. Last, we develop an analytical combinatorial approach to compute the frequency of best reply cycles in the case of uncorrelated payoffs.

RESULTS

Best reply structure

We begin by introducing a framework that we will demonstrate provides a useful estimate of the likelihood that the family of learning algorithms we analyze will converge to a fixed point. As we will demonstrate, this provides a kind of skeleton that can be analyzed to give a first-order approximation to the stability problems the algorithm will encounter as the players try to learn the game. The terminology that we will introduce is summarized in Table 1.

Assume a two-player normal-form game in which the players are Row and Column, each playing actions (or moves, or pure strategies) $i, j = 1, \ldots, N$. A best reply is the action that gives the best payoff in response to a given action by an opponent. The best reply structure is the arrangement of the best replies in the payoff bimatrix, i.e., the two matrices describing both players' payoffs. (In this paper, we will use the term "payoff matrix" to mean the bimatrix.) Under best reply dynamics, each player myopically responds with the best reply to the opponent's last action. We consider a particular version of best reply dynamics in which the two players alternate moves, each choosing her best response to her opponent's last action.

To see the basic idea, consider the game with $N = 4$ shown in Fig. 1A. Suppose we choose (1, 1) as the initial condition. Assume that Column moves first, choosing action $S^C = 2$, which is the best response to Row's action $S^R = 1$. Then, Row's best response is $S^R = 2$, then Column moves $S^C = 1$, etc. This traps the players in the cycle $(1, 1) \rightarrow (1, 2) \rightarrow (2, 2) \rightarrow (2, 1) \rightarrow (1, 1)$, corresponding to the red arrows. We call this a best reply 2-cycle, because each player moves twice. This cycle is an attractor, as can be seen by the fact that starting at (3, 2) with a play by Row leads to the cycle. The first mover can be taken randomly; if the players are on a cycle, this makes no difference, because at most one player has incentive to deviate from the current situation. However, when off an attractor, the order of the moves can be important. In this example, for instance, there are two attractors: If we begin at (3, 2) with a play by Column, we will arrive in one step at the best reply fixed point at (3, 3) (shown in blue). A fixed point of the best reply dynamics is a pure strategy Nash equilibrium (NE).

We characterize the set of attractors of best reply dynamics in a given $N \times N$ payoff matrix Π by a best reply vector $\boldsymbol{\upsilon}(\Pi) = (n_N, \ldots, n_2, n_1)$, where n_1 is the number of fixed points, n_2 is the number of 2-cycles, etc. For instance, $\upsilon = (0, 0, 1, 1)$ for the example in Fig. 1.

It is useful to reduce the payoff matrix to a best reply (bi-)matrix as shown in Fig. 1B. This is done by replacing all best replies for each player by one and all other entries by zero. The best reply matrix has the same best reply structure as the payoff matrix it is derived from, but it ignores any other aspect of the payoffs. The best reply matrix has the same cycles and pure strategy NE as the original game, but in general, the mixed strategy NE differ. Once the attractors of the best reply dynamics are known, it is trivial to list all the mixed strategy NE. There is one mixed NE associated with each best reply cycle, and there is one associated with all possible combinations of the cycles and fixed points. The mixed strategy NE associated with a given cycle corresponds to randomly playing each action with the frequency with which it is visited on the cycle. For example, the mixed strategy equilibrium associated with the best reply cycle in Fig. 1B is $\mathbf{x}, \mathbf{y} = (0.5, 0.5, 0, 0), (0.5, 0.5, 0, 0)$. The mixed strategy NE associated with each possible combination of cycles and fixed points corresponds to playing the average over the combined action sets. For example, in Fig. 1B, the mixed NE associated with the combination of the cycle and the fixed point is $\mathbf{x}, \mathbf{y} = (0.33, 0.33, 0.33, 0), (0.33, 0.33, 0.33, 0)$. For the best reply matrix, there are no other mixed strategy NE.

In moving from the best reply matrix to the original game, some of the mixed NE may survive and others may not, and new mixed NE may be introduced. For example, in Fig. 1A, none of them survive, and there are two mixed equilibria at $\mathbf{x}, \mathbf{y} = (0.32, 0, 0, 0.68), (0.36, 0.64, 0, 0)$ and at $\mathbf{x}, \mathbf{y} = (0.26, 0.15, 0, 0.59), (0.32, 0.24, 0.44, 0)$, which have no relation to those of the associated best reply dynamics. We make these statements quantitative for an ensemble of 1000 randomly generated games with $N = 10$ in section S2.

Learning dynamics

To address the question of when learning converges, we have studied six different learning algorithms. These are chosen to span different information conditions and levels of rationality. Our focus is on algorithms that have empirical support. This includes algorithms that are

Table 1. Terminology. NE, Nash equlibrium.

Best reply	Action that gives the best payoff in response to a given action by an opponent
Best reply structure	Arrangement of the best replies in the payoff matrix
Best reply matrix	Derived payoff matrix, with one for the best reply to each possible move of the opponent and zero everywhere else
Best reply dynamics	Simple learning algorithm in which the players myopically choose the best reply to the last action of their opponent
Best reply *k*-cycle	Closed loop of best replies of length *k* (each player moves *k* times)
Best reply fixed point	Pure NE, i.e., the action for each player that is a best reply to the move of the other player
Best reply vector υ	List of the number of distinct attractors of the best reply dynamics, ordered from longest cycles to fixed points
Free action/free best reply	Best reply to an action that is neither part of a cycle nor a fixed point
Best reply configuration	Unique set of best replies by both players to all actions of their opponent

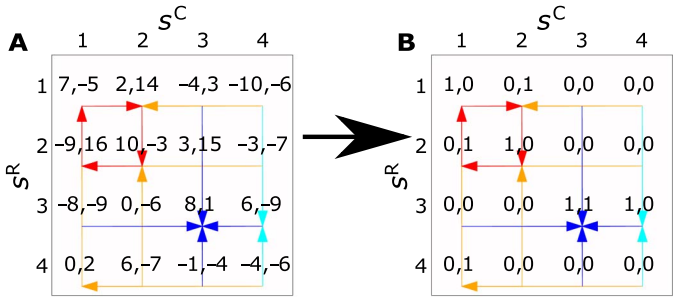

Fig. 1. Illustration of the best reply structure. $S^R = \{1, 2, 3, 4\}$ and $S^C = \{1, 2, 3, 4\}$ are the possible actions of players Row and Column, and each cell in the matrix represents their payoffs (Row is given first). The best response arrows point to the cell corresponding to the best reply. The vertical arrows correspond to player Row, and the horizontal arrows correspond to player Column. The arrows are colored red if they are part of a cycle, orange if they are not part of a cycle but lead to one, blue if they lead directly to a fixed point, and cyan if they lead to a fixed point in more than one step. The best reply matrix in (**B**) is a Boolean reduction that is constructed to have the same best reply structure as the payoff matrix in (**A**), but to only have one and zero as its entries.

used in biology for purposes such as animal learning or population genetics as well as those that are used to fit experiments about human behavior in playing games in laboratory settings. We provide a short summary of each of the six algorithms in Materials and Methods and present details in section S1. Here, we simply discuss the basic idea behind each algorithm and give our reasons for including it in this study.

Reinforcement learning (*36*) is based on the idea that players are more likely to play actions that yielded a better payoff in the past. It is the standard learning algorithm that is used with limited information and/or without sophisticated reasoning, such as in animal learning. We study the Bush-Mosteller implementation, which has been extensively tested in experiments (*37*).

Fictitious play (*38*, *39*) requires more sophistication, as it assumes that the players construct a mental model of their opponent. Each player assumes that the empirical distribution of her opponent's past actions is her mixed strategy, and plays the best response to this belief. A commonly used variant—weighted fictitious play—assumes discounting of past actions. The empirical evidence for standard versus weighted fictitious play is mixed (*40*). From a theoretical point of view, standard fictitious play converges to mixed strategy NE in many cases, while weighted fictitious play cannot do so (*41*). As we will see, our best reply formalism does not work as well if learning algorithms frequently reach mixed equilibria. We choose the standard version of fictitious play to illustrate this point, as it provides a stronger test.

Replicator dynamics (*42*) is commonly used in ecology and population biology. It represents the evolution of certain traits in a population over generations, with the fitness of each trait depending on the shares of all other traits in the population. Replicator dynamics can also be viewed as a learning algorithm in which each trait corresponds to an action (*43*). Here, we consider two-population replicator dynamics and not the more standard one-population version because, in the one-population version, the payoff matrices of the two players are, by definition, symmetric, whereas we want to study the dependence on the correlation of the payoffs.

Experience-weighted attraction (EWA) has been proposed (*44*) to generalize reinforcement learning and fictitious play and has been shown to fit experimental data well. The key of its success is a parameter that weighs realized payoffs versus forgone payoffs. EWA also includes other parameters such as memory and payoff sensitivity. In contrast to the three learning algorithms above, its parameters are crucial in determining convergence: For some parameter values, it always converges to fixed points that can be arbitrarily far from Nash or correlated equilibria (*45*). As we discuss at length in the Supplementary Materials, lacking experimental guidance for parameters in generic games, we choose parameter values that potentially allow convergence close to NE. [EWA does not converge close to more general correlated equilibria unless they correspond to NE (*45*).]

The above algorithms are all based on the concept of batch learning, which is convenient because it means that they are deterministic. Under batch learning, the players observe the actions of their opponent a large number of times before updating their strategies and so learn based on the actual mixed strategy of their opponent at each point in time. The deterministic assumption is useful to identify stationary states numerically. In many experimental situations, it is more realistic to assume that players update their strategies after observing a single action by their opponent, which is randomly sampled from her mixed strategy. This is called online learning. As an example of online learning, we study the stochastic version of EWA.

The above five algorithms are all backward looking. To compare to an example with forward-looking behavior, we use level-*k* learning (*46*), or anticipatory learning (*47*), in which the players try to outsmart their opponent by thinking *k* steps ahead. There is considerable empirical support for the idea of level-*k* thinking (*48*), and some studies for anticipatory learning specifically (*49*). Here, we implement level-*k* reasoning by using level-2 EWA learning. Both players assume that their opponent is a level 1 learner and update their strategies using EWA. So the players try to preempt their opponent based on her predicted action, as opposed to acting based solely on the frequency of her historical actions. This provides a measure of forward-looking behavior.

To summarize, we have chosen these six algorithms because they are important in their own right and because they are illustrative of different properties. We are interested in algorithms that are used in ecology and biology or that display the kind of bounded rationality that is observed in laboratory experiments with human subjects (*36*, *37*, *40*, *44*, *49*). We are specifically not studying sophisticated algorithms that are relevant for computer science or mechanism design but are not relevant from a behavioral point of view. These algorithms may have very different convergence properties from those studied here. An example of the kind of algorithm we are not studying is regret matching (*17*), which is considered an "adaptive heuristic" and is boundedly rational and myopic. However, it still requires that players compute what payoff they would have received had they played any other action in all previous time steps, and there is no empirical evidence that players in human experiments use it. The other algorithms that reach correlated equilibria, such as calibrated learning, are even more sophisticated.

How does the best reply structure shape learning dynamics?

Our working hypothesis is that the best reply structure influences the convergence properties of these learning algorithms, even if the learning trajectories may not follow best reply cycles in detail. More specifically, our hypothesis is that the presence or absence of best reply cycles is correlated with the stability of the learning dynamics. Learning is more likely to converge when there are no best reply cycles and less likely to converge when they are present. We cannot prove this analytically in generic games, but it is supported by anecdotal examples. It is impossible to pres-

ent these in detail, but here, we present a few representative samples to motivate the correspondence. In the next section, we present more quantitative evidence.

To develop intuition into whether and why the best reply structure could predict convergence of the learning algorithms above, in Fig. 2, we analyze four games with $N = 3$, showing some learning trajectories in a three-dimensional projection of the four-dimensional strategy space (there are six components of the mixed strategy vectors with two normalization constraints). The axis labels x_1 and x_2 are the probabilities of player Row to play actions $s^R = 1$ and $s^R = 2$, respectively, and y_1 is the probability for Column to play $s^C = 1$. The corners of the strategy space $(x_1, y_1, x_2) = (1, 1, 0)$ and $(0, 1, 1)$ correspond to the action profiles (1, 1) and (2, 1), respectively.

In Fig. 2A, we consider a best reply matrix with a 2-cycle and a single pure strategy NE. Our first illustration uses replicator dynamics. The attractors of the learning algorithm closely mimic the attractors of the best reply dynamics; all trajectories go to either the fixed point or the limit cycle, depending on the initial condition. The limit cycle in this example corresponds to the best reply cycle, as we always have $y_3 = 0$. (Coordinate y_3 is not shown due to the three-dimensional projection of the six-dimensional system.) Reinforcement learning, EWA, and EWA with noise behave similarly. In contrast, fictitious play always converges to a fixed point, either to the pure strategy NE or to the mixed strategy equilibrium in the support of the cycle, depending on initial conditions. We never observe it converging to the other mixed NE, which corresponds to the combination of the two attractors. In section S2, we show that in generic games with $N = 10$, fictitious play is more likely to converge to mixed equilibria in the support of a best reply cycle than to other mixed equilibria, with respect to the proportions of existing mixed equilibria. Level-k EWA also converges (close) to the same mixed equilibrium. As we will show quantitatively in the following section, level-k EWA behaves like fictitious play in games with few actions and more like the other four algorithms in games with many actions. For best reply matrices, it is not surprising that the learning dynamics mimic the best reply structure, but because the learning algorithms have free parameters and memory of the past, it is not obvious that they should do so this closely.

The payoff matrix in Fig. 2B has the same best reply structure as that in Fig. 2A but has generic payoffs. To show a wider variety of learning algorithms, we illustrate the learning dynamics this time with reinforcement learning and EWA with noise. In both cases, the learning trajectories either converge to the pure strategy NE or do not converge. Reinforcement learning converges to a limit cycle that is related to the best reply cycle, even if the trajectory is distorted and penetrates more deeply into the center of the strategy space, and there is a similar correspondence for EWA with noise. As in Fig. 2A, replicator dynamics and EWA behave similarly, while fictitious play and level-k EWA converge to or close to mixed equilibria.

To see what happens once the best reply cycle is removed, in Fig. 2C, we consider the same payoff matrix in Fig. 2B, except that the payoff for Row at (3, 1) is 2.0 instead of 0.1. In this case, the only attractor of the best reply dynamics is the pure strategy NE at (3, 3). For

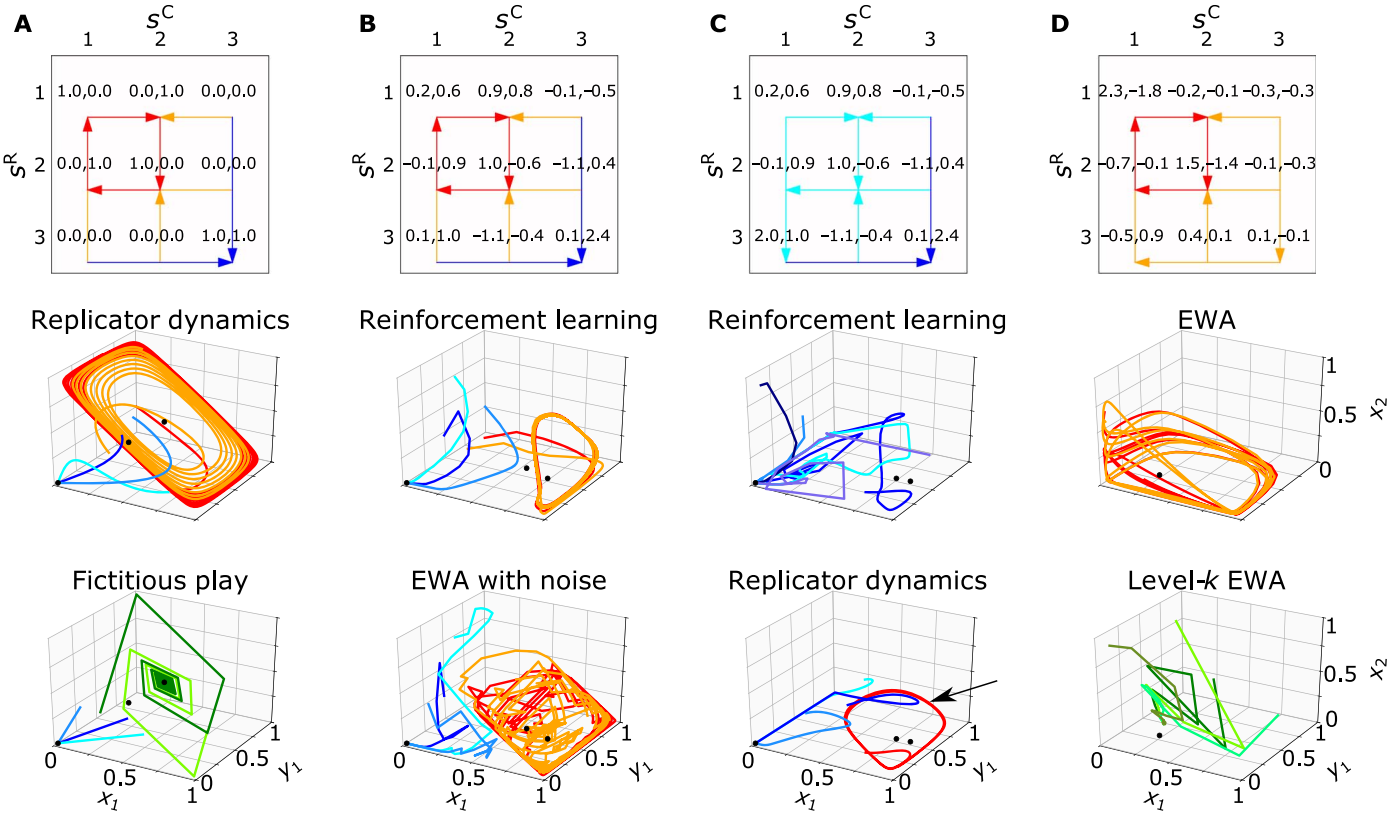

Fig. 2. How the best reply structure influences convergence of six learning algorithms. For each payoff matrix (**A**) to (**D**), we show the dynamics of two learning algorithms. We plot the probabilities x_1 and x_2 for Row to play actions $s^R = 1$ and $s^R = 2$, respectively, and the probability y_1 for Column to play $s^C = 1$. The black dots are the positions of the NE: For payoff matrices (A) to (C), the dot at $x_1 = x_2 = y_1 = 0$ is the pure strategy NE; all other dots are mixed strategy equilibria. The trajectories are colored in shades of blue if they converge to a pure equilibrium, in shades of green if they converge to a mixed equilibrium, and in shades of red if they do not converge. NE, Nash equilibrium.

all initial conditions, reinforcement learning converges to the pure strategy NE, illustrating how removing the best reply cycle makes the fixed point globally stable. However, for replicator dynamics, there exist some initial conditions leading instead to a limit cycle, suggesting that a small basin of attraction for unstable learning dynamics may still exist even in the absence of best reply cycles. (Fictitious play and level-k EWA behave like reinforcement learning, while EWA and EWA with noise behave like replicator dynamics.)

This case illustrates an important example of how the learning dynamics can qualitatively deviate from the best reply dynamics. When the replicator dynamics reaches the point $(x_1, y_1, x_2) = (0.86, 0.87, 0.14)$, indicated by a black arrow in the figure, the expected payoff for Row is 0.25. If Row switches to $s^R = 3$ (his best reply), his payoff raises to 1.60. If instead Row switches to $s^R = 1$, Row's expected payoff still increases to 0.29, so we say that action 1 is a better reply. With replicator dynamics, the probability for actions corresponding to better replies grows over time, although at a lower rate than best replies. Here, because of past play, x_3 is very small (of the order of 10^{-53}), so x_1 increases to 1 before x_3 can reach a similar value, and the dynamics gets trapped in the cycle shown in the figure.

In Fig. 2D, we consider a payoff matrix with a 2-cycle and no pure strategy NE. EWA reaches a chaotic attractor for all initial conditions. Note that this attractor shares some of the structure of the best reply cycle, visiting the two corners of the strategy space corresponding to $(x_1, y_1, x_2) = (1, 0, 0)$ and $(1, 1, 0)$ closely and the other two more loosely. Level-k EWA, in contrast, converges close to the mixed NE in the center of the chaotic attractor. [Due to finite memory and finite payoff sensitivity, EWA, EWA with noise, and level-k EWA cannot reach mixed equilibria exactly (*45*).] The other algorithms behave similarly: Both EWA with noise and replicator dynamics reach a chaotic attractor, while reinforcement learning follows a limit cycle. Fictitious play fluctuates close to the mixed equilibrium, with each component of the mixed strategy vector at a maximum distance of 0.03 from the equilibrium point.

We thus see that, despite the fact that the learning dynamics of these algorithms do not mimic the best reply dynamics in detail, in most cases remnants of it are seen in the learning dynamics. Even if the correspondence is not exact, the best reply structure influences the learning dynamics.

Quantitative evidence

We now quantitatively test the hypothesis that the presence of best reply cycles is positively correlated to nonconvergence, at least for these learning algorithms. To do this, we first prove a very useful theorem relating the configuration of cycles and fixed points to the relative size of the basin of attraction for unstable best reply dynamics. Recalling that n_k is the number of best reply cycles of length k, let $C = \sum_{k=2}^{N} n_k k$ be the number of actions that are part of best reply cycles. The quantity

$$\mathcal{F}(\upsilon) = C/(C + n_1) \tag{1}$$

measures the relative number of actions on attractors that are part of cycles versus fixed points. For the example of Fig. 1, where there is one fixed point and one cycle, $\mathcal{F}(0, 0, 1, 1) = 2/3$.

This also measures the relative size of the basins of attraction of cycles versus fixed points under best reply dynamics. This is not true for a given individual game, but as we show in section S3.1, it is true for the average over the ensemble of all games with a given best reply vector υ. For example, for the game shown in Fig. 1, five of eight of the initial conditions lead to the cycle and three of eight lead to the fixed point. However, if one averages over all possible 4×4 games that have one best reply cycle of length 2 and a single pure strategy NE and no other best reply attractors (giving an equal weighting to each possible game), the relative size of initial conditions leading to the best reply cycle is exactly 2/3. As we show in the Supplementary Materials, this is true in general.

To test the relationship between best reply dynamics and the convergence properties of our family of learning algorithms, we generate games at random. This is done by populating each payoff bimatrix with $2N^2$ normally distributed numbers; as discussed in section S1.2, the normal distribution is the maximum entropy distribution and so is the natural choice. Here, we let the payoffs of the two players be uncorrelated. We then simulate the learning process of the players in a repeated game, holding the payoff matrix fixed. Convergence to pure and mixed strategy NE is checked using the criteria explained in Materials and Methods and in the Supplementary Materials. We generate 1000 different games; for each game, we simulate all six learning algorithms starting from 100 different initial conditions. As a point of comparison, we repeat the entire process using the best reply matrices associated with each of the 1000 randomly generated games. Results with $N = 20$ are reported in Fig. 3, and results with $N = 5$ and $N = 50$ are given in figs. S8 and S9.

Figure 3 compares the share of best reply cycles $\mathcal{F}(\upsilon)$ to the nonconvergence frequency. To test the relationship between the best reply vector υ and the learning dynamics, we group together the results for payoff matrices with the same υ and plot a circle whose radius is proportional to the logarithm of the number of times this was sampled. We place each best reply vector on the horizontal axis according to its share of best reply cycles $\mathcal{F}(\upsilon)$. On the vertical axis, we plot the average frequency of nonconvergence for results with this best reply vector. Thus, if the best reply structure perfectly predicts the rate of convergence of the other learning algorithms, all circles should be centered on the identity line. We estimate the weighted correlation coefficient R^2_w using weights corresponding to the number of times each best reply vector was sampled.

The results when we simulate using the best reply matrices are shown in the top row of Fig. 3. The correlation is nearly one for all the algorithms except fictitious play. On one hand, given that best reply matrices do not have better replies, this may not seem too surprising. But on the other hand, these learning algorithms all use memory of the past (beyond the most recent action) and most of them have free parameters, while best reply dynamics has neither of these. Given that, it is remarkable that the size of the basins of attraction of learning algorithms with different functional forms is so well approximated by best reply dynamics. The failure for fictitious play is due to its strong tendency to converge to mixed strategy NE, although even in this case it typically fails to converge for longer cycles. [Fictitious play converges in most cases in the presence of 2-cycles but fails to converge with 3-cycles, in line with Miyazawa (*50*) and Shapley (*51*).] The results in the top panels of Fig. 3 reinforce the intuition that, for best reply matrices, the attractors of the learning algorithms closely match the best reply dynamics for all algorithms except fictitious play.

The general case is shown in the bottom row of Fig. 3. The correlations are still very strong, with weighted correlation coefficient $R^2_w = 0.78$ for fictitious play and $R^2_w \approx 0.84$ for all of the other algorithms. The best reply dynamics are not a perfect predictor of convergence, but the predictions are still good enough to be very useful.

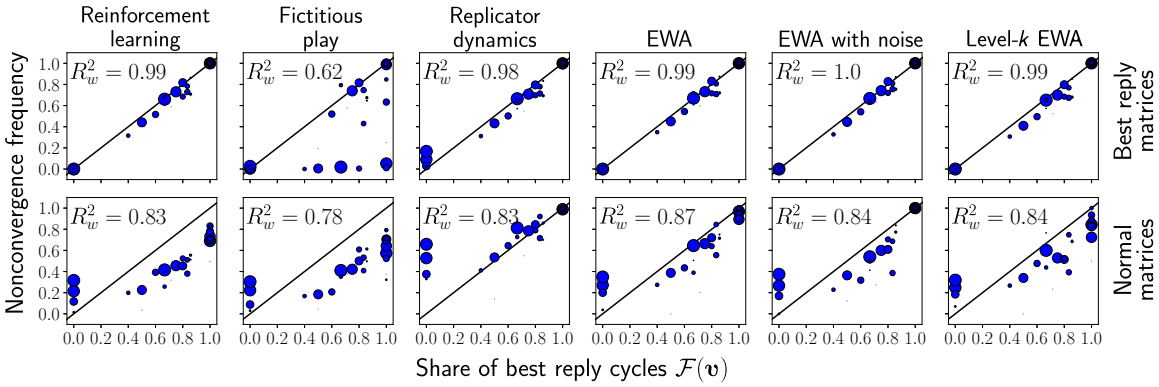

Fig. 3. Test for how well the best reply structure predicts nonconvergence for six learning algorithms. We generate 1000 random payoff matrices with $N = 20$ actions, and for each of these, we simulate learning 100 times based on random initial conditions. The same process is repeated using the best reply matrix associated with each of the 1000 random games. Each circle corresponds to a specific best reply vector υ. Its size is the logarithm of the number of times a payoff matrix with υ was sampled. The horizontal axis is the share of best reply cycles $\mathcal{F}(\upsilon)$. For example, the largest circle at $\mathcal{F}(\upsilon) = 0.66$ corresponds to $\upsilon = (0, \ldots, 0, 1, 1)$. The vertical axis gives the frequency of nonconvergence in the simulations, averaged over all payoff matrices and initial conditions having the same υ. The top row shows results for the best reply matrices, and the bottom row shows results using normally distributed payoffs. The identity line is plotted for reference.

This analysis also gives clues about some of the other factors that cause nonconvergence. For example, even when $\mathcal{F}(\upsilon) = 0$, indicating that best reply cycles are absent, convergence is not certain. This is evident from the vertical column of circles on the left of each figure. This column corresponds to best reply vectors with no cycles, i.e., those of the form $\upsilon = (0, \ldots, 0, 0, x)$, where $x = 1, 2, 3, 4$ is the number of distinct fixed points. The highest circle corresponds to a single fixed point, the one below it to two fixed points, etc. In the case where there is a unique pure NE and no cycles, the nonconvergence frequency is typically around 35%, dropping to about 20% if there are two pure equilibria. The presence of multiple pure equilibria makes the existence of better reply cycles like the one in Fig. 2C less likely. Conversely, there are also cases of convergence when there are cycles without any fixed points. This corresponds to the column of vertical circles on the right, with $\mathcal{F}(\upsilon) = 1$, for level-$k$ EWA, fictitious play, and reinforcement learning.

Nonconvergence in the presence of a unique pure NE with no cycles is particularly pronounced for replicator dynamics, where nonconvergence happens about 60% of the time when there is a single fixed point. As we explain in the Supplementary Materials, this is in part due to numerical limitations that become more serious as N grows, so we will drop observations of replicator dynamics if $N \geq 50$.

Convergence in the absence of fixed points is entirely due to convergence to mixed strategy NE. As demonstrated in fig. S10, this effect is almost independent of $\mathcal{F}(\upsilon)$ and so causes a constant offset. This is the reason that, for $\mathcal{F}(\upsilon) > 0$, the circles are below the identity line. This effect is particularly strong for reinforcement learning, where convergence to mixed equilibria occurs 21% of the time, and fictitious play, for which it occurs 32% of the time. In contrast, this effect is absent for (two-population) replicator dynamics, which cannot converge to mixed strategy NE (*52*). For the other algorithms, the frequency of convergence to mixed strategy NE is less than 15%. Once these effects are taken into account, there is a strong proportionality between $\mathcal{F}(\upsilon)$ and the nonconvergence rate for all of the learning algorithms, including fictitious play.

In fig. S11, we show the correlation matrix of the convergence of the six learning algorithms. We find that convergence co-occurs on average 60% of the time, suggesting a significant degree of heterogeneity among the algorithms, in line with the intuition in Fig. 2.

In summary, even if $\mathcal{F}(\upsilon)$ ignores better replies and underestimates the convergence to mixed strategy NE, a robust correlation between the average probability of convergence and the share of best reply cycles exists. This indicates that the best reply structure and the stability of these algorithms are closely linked.

Variation of the best reply structure

We now investigate the prevalence of best reply cycles and fixed points as we vary the properties of the games, and test the prediction of nonconvergence as we vary parameters. The parameters that we investigate are the number of possible actions N and the correlation Γ between the payoffs of the two players. (Later, we also consider a parameter ξ to interpolate between random and potential games.)

As before, we generate games at random, but we now impose constraints on the ensemble from which the games are drawn. To understand how Γ affects convergence, we generate payoff matrices by drawing from a bivariate normal distribution so that the expected value of the product of the payoffs to players Row and Column for any given combination of actions is equal to Γ. A negative correlation, $\Gamma < 0$, implies that the game is competitive, because what is good for one player is likely to be bad for the other. The extreme case is where $\Gamma = -1$, meaning the game is zero sum. In contrast, $\Gamma > 0$ encourages cooperation, in the sense that payoffs tend to be either good for both players or bad for both players.

In Fig. 4, we show how the share of best reply cycles varies with N and Γ. For a given value of N and Γ, we randomly generate payoff matrices and compute the average share of best reply cycles $\langle \mathcal{F}(\upsilon) \rangle_{N,\Gamma}$. We compare $\langle \mathcal{F}(\upsilon) \rangle_{N,\Gamma}$ to the average frequency of nonconvergence of the six learning algorithms. The agreement is generally good. The striking exception is fictitious play, which has a high convergence rate when N is large and Γ is negative. This is at odds with the other algorithms, which rarely converge in this parameter range and agree closely with the prediction. Thus, as expected, fictitious play must be regarded somewhat differently than the other algorithms.

We would like to emphasize that these predictions are made a priori—they do not involve fitting any parameters. With the exception of fictitious play, the predictions fit the overall trends in the data, both estimating the magnitude correctly and capturing the functional

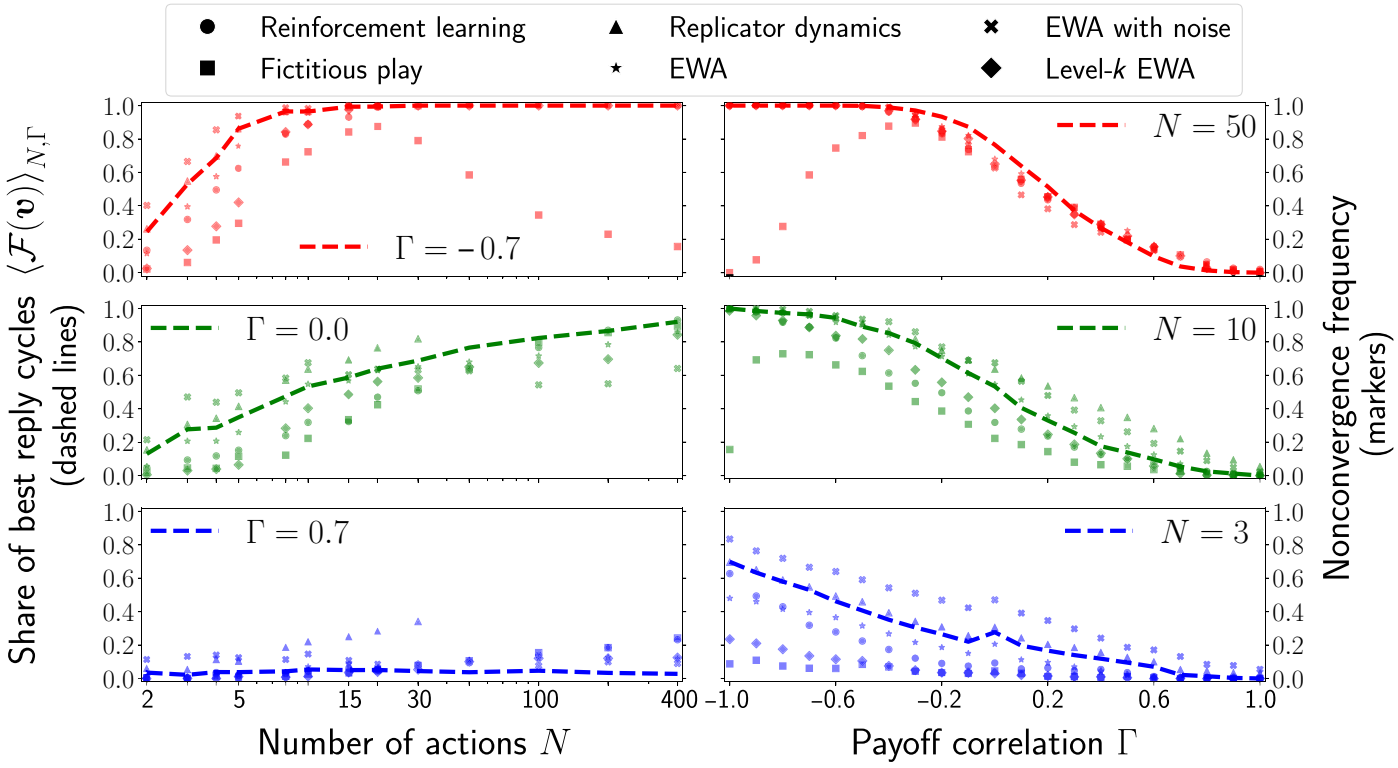

Fig. 4. How the share of best reply cycles predicts convergence as a function of the number of actions *N* and the competition parameter Γ. Dashed lines are the average share of best reply cycles $\langle \mathcal{F}(\upsilon) \rangle_{N,\Gamma}$ for those values of *N* and Γ. Markers are the fraction of simulation runs in which the learning algorithms do not converge.

dependence. For the other five algorithms, the predictions are particularly good when *N* is large and Γ is strongly negative, with a near-perfect fit for $N > 10$.

What are the conclusions about the circumstances in which convergence is likely as we vary *N* and Γ? When Γ is positive (meaning that the game is not competitive), convergence is almost guaranteed, regardless of *N*, but when Γ is strongly negative, convergence is likely only when *N* is very small. At $\Gamma = -0.7$, for $N = 4$, the nonconvergence rate is roughly 70%, quickly rising to approach 100% by $N = 8$. Our analysis indicates that this is because complicated competitive games are dominated by cycles. In this region of the parameter space, acyclic games are extremely rare. Therefore, dominance-solvable, coordination, potential, and supermodular games are atypical.

Last, we sketch how one might introduce other constraints to study deviations from the classes of games above. Here, we just give a high-level discussion, reporting the technical details in section S4. We consider potential games, in which all differences in payoffs for unilaterally switching to another action can be expressed using a global potential function (*8*). The potential function maps each action profile to a real number. We generate payoff matrices at random, and for each payoff matrix, we also generate an associated potential function. We then modify the game so that the differences in payoffs conform to the potential function. This is tuned by a parameter $\xi \in [0, 1]$: When $\xi = 0$, there is no modification, and so the game is completely random, and when $\xi = 1$, the payoff matrix describes a perfect potential game.

What is the effect of changing *N* and ξ? We repeat the process described above for 1000 payoff matrices for values $N = 3, 10, 50$ and $\xi \in [0.0, 0.1, \ldots, 0.5, \ldots, 0.9, 1.0]$. As we show in fig. S12, when ξ approaches 1, the share of best reply cycles $\langle \mathcal{F}(\upsilon) \rangle_{N,\xi}$ goes to zero as expected. However, it does so in a highly nonlinear way, particularly for large *N*. When $N = 3$ and $\xi = 0.8$, it is $\langle \mathcal{F}(\upsilon) \rangle_{N,\xi} = 0.05$ only, but for $N = 50$ and $\xi = 0.8$, it is $\langle \mathcal{F}(\upsilon) \rangle_{N,\xi} = 0.35$. This suggests that, in some situations, small deviations from commonly studied classes of games may cause a significant increase in the share of best reply cycles, making them less stable under learning.

Analytical approach

For $\Gamma = 0$, it is possible to derive analytically how the best reply structure varies with *N*. Our derivation follows the framework of statistical mechanics, suggesting that problems such as the existence of phase transitions in the best reply structure may be studied within our formalism. We define a best reply configuration as a unique set of best replies by both players to all possible actions of their opponent. Because the best reply matrix is a Boolean matrix, for any given *N*, there is a countable number of possibilities. The total number of possible best reply configurations is N^{2N}. For uncorrelated payoffs, $\Gamma = 0$, all best reply configurations are equally likely. Therefore, we can compute the frequency $\rho(\upsilon)$ for any set of attractors υ by counting the number of best reply configurations leading to υ. In the jargon of statistical mechanics, this means that we are assuming a microcanonical ensemble of games.

Here, we just sketch the derivation, referring the reader to section S5.1 for a detailed explanation. Because of independence, the frequency $\rho(\upsilon)$ can be written as a product of terms *f* corresponding to the number of ways to obtain each type of attractor multiplied by a term *g* for free actions (best replies that are not on attractors). We denote by *n* the number of actions per player, which are not already part of cycles or fixed points.

The function $f(n, k)$ counts the ways to have a k-cycle (including fixed points, which are cycles of length $k = 1$)

$$f(n,k) = \binom{n}{k}^2 k!(k-1)! \tag{2}$$

where the binomial coefficient means that, for each player, we can choose any k actions out of n to form cycles or fixed points, and the factorials quantify all combinations of best replies that yield cycles or fixed points with the selected k actions. For instance, in Fig. 1, for each player, we can choose any two actions out of four to form a 2-cycle, and for each of these, there are two possible cycles (one clockwise and the other counterclockwise). The number of ways to have a 2-cycle is $f(4, 2) = 72$. Similarly, for each player, we can select any action out of the remaining two to form a fixed point, in $f(2, 1) = 4$ ways.

In this example, for both players, we can still freely choose one best reply, provided that this does not form another fixed point (otherwise, the best reply vector would be different). In Fig. 1, the free best replies are (3, 4) for Row and (4, 1) for Column. In general, $g_N(n, d)$ counts the number of ways to combine the remaining n free best replies in a $N \times N$ payoff matrix so that they do not form other cycles or fixed points

$$g_N(n,d) = N^{2n} - \sum_{k=1}^{n} f(n,k) g_N(n-k, d+1)/(d+1) \tag{3}$$

The first term N^{2n} quantifies all possible combinations of the free best replies, and the summation counts the "forbidden" combinations, i.e., the ones that form cycles or fixed points. This term has a recursive structure. It counts the number of ways to form each type of attractor and then the number of ways not to have other attractors with the remaining $n - k$ actions. Note that N is a parameter and therefore is indicated as a subscript, while n is a recursion variable. d denotes the recursion depth. Last, the division by $d + 1$ is needed to prevent double, triple, etc., counting of attractors. In the example of Fig. 1, $g_4(1, 0) = 15$.

For any given best reply vector, $\upsilon = (n_N, \ldots, n_2, n_1)$, the general expression for its frequency ρ is

$$\rho(\upsilon) = \left(\prod_{k=1}^{N} \prod_{j=1}^{n_k} \frac{f\left(N - \sum_{l=k+1}^{N} n_l l - (j-1)k, k\right)}{j} \right) \times g_N\left(N - \sum_{l=1}^{N} n_l l, 0\right)/(N^{2N}) \tag{4}$$

The product in the first brackets counts all possible ways to have the set of attractors υ. The first argument of f, $N - \sum_{l=k+1}^{N} n_l l - (j-1)k$, iteratively quantifies the number of actions that are not already part of other attractors. The division by j, like the division by $d + 1$ in Eq. 3, is needed to prevent double, triple, etc., counting of attractors. The second term g_N counts all possible ways to position the free best replies so that they do not form other attractors. The first argument of g_N is the count of actions that are not part of attractors, and the initial recursion depth is 0. Last, we obtain the frequency by dividing by all possible configurations N^{2N}. For the payoff matrix in Fig. 1, $\rho(0, 0, 1, 1) = f(4, 2) f(2, 1) g_4(1, 0)/4^8 = 0.07$.

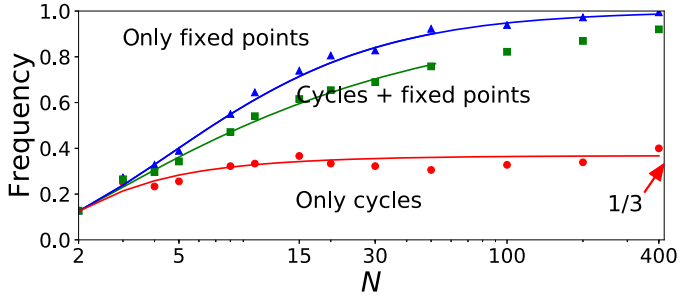

Fig. 5. Comparison of analytical predictions about best reply cycles to numerical simulations when $\Gamma = 0$. Markers are numerical results, and solid lines are analytical results. Red circles depict the frequency of randomly generated payoff matrices with no fixed points ($\mathcal{F}(\upsilon) = 1$), and blue triangles show the frequency with at least one cycle ($\mathcal{F}(\upsilon) > 0$). The text in the figure refers to the area delimited by solid lines, e.g., "Cycles + fixed points" means that the fraction of payoff matrices with both cycles and fixed points is the distance between the red and blue lines. Last, green squares represent the average share of best reply cycles $\mathcal{F}_N$; this is discontinued at $N = 50$ due to excessive computational cost (see section S5.2).

Equation 4 can then be used to compute the ensemble average of the share of best reply cycles $\mathcal{F}$ for any given N

$$\mathcal{F}_N = \sum_{\upsilon} \rho(\upsilon) \mathcal{F}(\upsilon) \tag{5}$$

summing over all possible υ s.t. $\sum_{k=1}^{N} n_k k \leq N$. It is also possible to calculate other quantities, including the fraction of payoff matrices without fixed points ($\mathcal{F}(\upsilon) = 1$) and without cycles ($\mathcal{F}(\upsilon) = 0$). We provide the expressions and explain their derivation in section S5.2.

In Fig. 5, we analyze the best reply structure for increasing values of N. We report, from bottom to top, the fraction of payoff matrices with no fixed points, the average share of best reply cycles $\mathcal{F}_N$, and the fraction of games with at least one cycle. For instance, for $N = 30$, 36% of the payoff matrices have no fixed points and 84% have at least one cycle (so 16% have no cycles and 48% have a mixture of cycles and fixed points), with an average $\mathcal{F}_N = 0.70$. There is a very good agreement between analytical results (solid lines) and Monte Carlo sampling (markers). The fraction of games with cycles is an increasing function of N; it is computationally intractable to compute this for large N, but it seems to be tending to one. However, the fraction of games with at least one fixed point seems to reach a fixed value for $N \to \infty$. In section S5.3, we show that this is approximated by 1/3, in agreement with numerical simulations and close to the exact result $1/e$ (*22*).

DISCUSSION

We have characterized instability in two-player, normal-form, generic games, showing that the best reply structure predicts the convergence frequency of a wide variety of learning algorithms. This result is remarkable because these algorithms have no explicit relationship to best reply dynamics. Best reply dynamics depends only on the other player's previous move, whereas these algorithms all have longer memory, and also because best reply dynamics has no free parameters, whereas most of these algorithms do. Why does this correspondence work? We conjecture that the presence of best reply cycles makes the existence of basins of attraction for unstable dynamics statistically more likely, while their absence makes it probable that pure strategy NE are globally stable fixed points. This is not always true—some learning dynamics may be

trapped in better reply cycles, and others sometimes converge to mixed strategy NE—but it is a valid approximation that enormously simplifies the problem. It makes it possible to use combinatorics to analytically explore the space of generic games under the microcanonical ensemble using the conceptual framework of statistical mechanics.

Why are the predictions of the best reply formalism this good? To understand this, one could explicitly treat the best reply payoff matrix as a first-order approximation and then study the behavior as one moves along a continuous path through the family of payoff matrices it is associated with. One could use this method to study the factors that cause deviations, such as the appearance and disappearance of mixed equilibria or better replies. This could potentially lead to a kind of perturbation expansion for understanding the convergence of learning algorithms in normal-form games. This is beyond the scope of this paper but could be an interesting topic of further study.

We have also shown that in the absence of any constraints on the payoffs, increasing the number of available actions per player makes best reply cycles dominant, thereby making convergence to equilibrium unlikely. This is akin to May's result (*14*) on large ecosystems. We considered a competition parameter that constrains the pairs of payoffs to both players. With negative correlation, the game is competitive (zero sum in the extreme case of perfectly anticorrelated payoffs), and best reply cycles become even more prevalent. Positive correlation instead makes best reply cycles rare.

We have also constrained the games so that the differences in payoffs conform to a global potential function, as in potential games. The correspondence is tuned by a parameter that interpolates between perfect random games and perfect potential games. We have shown that small deviations from perfect potential games entail a substantial share of best reply cycles when the number of actions is large.

Many other constraints could be included. Supermodular games have also received great attention in the literature (*11*). In these games, actions are ordered, and if one player increases her action, all other players' marginal payoffs increase. (This is the concept of strategic complementarities.) It would be possible to generate games at random with this constraint and to see how the share of best reply cycles varies as perfect supermodular games are approached. In general, our formalism makes it possible to evaluate whether constraints make ensembles of normal-form games more or less stable. In most situations, for the learning algorithms we study here, it is enough to check if the constraints make best reply cycles common or rare, without the need to simulate learning dynamics.

We have studied normal-form games here because they are tractable, but our study can potentially be extended to other types of games. For example, the sequencing of different actions and the existence of private information are properly modeled in extensive form games. While our theory does not apply directly, extensive form games have a normal-form representation (*1*), suggesting that such an extension should be possible. Similarly, we have already begun the study of games with more than two players. A previous study of competitive games with an infinite number of actions N suggests that nonconvergence becomes even more likely (*53*), and our preliminary results suggest that this is also true for $N < \infty$.

In ecology, the mere fact that ecosystems are so persistent makes it clear from the outset that they must be fairly stable. In contrast, there are many biological and social systems that fluctuate in time, and it is not clear a priori whether their dynamics are exogenous or endogenous. When these systems are modeled using game theory, there is no prior that says that they should be stable. Thus, unlike ecosystems, where the answer is known in advance, in these systems, we should be open-minded about whether they are generically stable or unstable. In the absence of selection mechanisms or intentionally designed stability, our approach sheds light on this question. In the presence of either of these, it provides a null hypothesis against which the effectiveness of these mechanisms can be measured.

These results are useful because they give a warning about situations in which the assumption of equilibrium is dangerous. There are many real-world situations where the number of possible actions is large and where payoffs are likely to be anticorrelated. In the absence of other constraints, our results suggest that, in these circumstances, equilibrium is unlikely to be a good behavioral assumption. Although equilibria exist, insofar as normal-form games apply, and insofar as the type of learning algorithms we have studied here is relevant, in these circumstances, convergence is unlikely.

MATERIALS AND METHODS

We summarize here the protocol that was used to simulate the learning algorithms in Figs. 3 and 4. We only report the minimal information that would allow replication of the results. A more detailed description, in which we provide behavioral explanations and mention alternative specifications, is given in section S1. We had to make arbitrary choices about convergence criteria and parameter values, but when testing alternative specifications, we found that the correlation coefficients had changed by no more than a few decimal units. This confirms a robust correlation between the share of best reply cycles and the nonconvergence frequency of the six learning algorithms.

Consider a two-player, N-actions normal-form game. We index the players by $\mu \in \{\text{Row} = R, \quad \text{Column} = C\}$ and their actions by i, $j = 1, \ldots N$. Let $x_i^\mu(t)$ be the probability for player μ to play action i at time t, i.e., the ith component of her mixed strategy vector. For notational convenience, we also denote by $x_i(t)$ the probability for player R to play action i at time t, and by $y_j(t)$ the probability for player C to play action j at time t. We further denote by $s_\mu(t)$ the action that is actually taken by player μ at time t, and by $s_{-\mu}(t)$ the action taken by her opponent. The payoff matrix for player μ is Π^μ, with $\Pi^\mu(i, j)$ as the payoff μ receives if she plays action i and the other player chooses action j. So if player Row plays action i and player Column plays action j, they receive payoffs $\Pi^R(i, j)$ and $\Pi^C(j, i)$, respectively.

Reinforcement learning

We only describe player Row, because the learning algorithm for Column is equivalent. Player Row at time t has a level of aspiration $A^R(t)$ that updates as

$$A^R(t+1) = (1-\alpha)A^R(t) + \alpha \sum_{i,j} x_i(t)\Pi^R(i,j)y_j(t) \qquad (6)$$

where α is a parameter. For each action i and at each time t, player Row has a level of satisfaction $\sigma_i^R(t)$ given by

$$\sigma_i^R(t) = \frac{\sum_{ij} x_i(t)y_j(t)(\Pi^R(i,j) - A^R(t))}{\max_{i,j}|\Pi^R(i,j) - A^R(t)|} \qquad (7)$$

All components of the mixed strategy vector are updated. The update rule is

$$x_i(t+1) = x_i(t) + x_i(t)\Delta x_i(t) + \sum_{j\neq i} x_j(t)\Delta x_{ij}(t) \tag{8}$$

Here, $\Delta x_i(t)$ is the contribution due to the choice of action i by player Row (which occurs with probability $x_i(t)$, hence the multiplying term), and $\Delta x_{ij}(t)$ is the contribution on action i due to the choice of another action j (i.e., a normalization update), each occurring with probability $x_j(t)$. We have

$$\Delta x_i(t) = \begin{cases} \beta\sigma_i^{\mathrm{R}}(t)(1-x_i(t)), & \sigma_i^{\mathrm{R}}(t) > 0 \\ \beta\sigma_i^{\mathrm{R}}(t)x_i(t), & \sigma_i^{\mathrm{R}}(t) < 0 \end{cases} \tag{9}$$

and

$$\Delta x_{ij}(t) = \begin{cases} -\beta\sigma_j^{\mathrm{R}}(t)x_i(t), & \sigma_j^{\mathrm{R}}(t) > 0 \\ -\beta\sigma_j^{\mathrm{R}}(t)\dfrac{x_j(t)x_i(t)}{1-x_j(t)}, & \sigma_j^{\mathrm{R}}(t) < 0 \end{cases} \tag{10}$$

with β being a parameter.

Starting from random mixed strategy vectors—the initialization of the mixed strategies will be identical for all learning algorithms that follow—and null levels of aspiration and satisfaction, we iterated the dynamics in Eqs. 6 to 10 for 5000 time steps (we set α = 0.2 and β = 0.5). To identify the simulation run as convergent, we only considered the last 20% of the time steps and the components of the mixed strategy vectors played with average probability greater than 0.05 in this time interval. If the standard deviation averaged over these components and time steps was larger than 0.01, the simulation run was identified as nonconvergent.

Fictitious play

Player Row calculates the jth component of the expected mixed strategy of Column at time T, which we denote by $\tilde{y}_j(T)$, as the fraction of times that j has been played in the past

$$\tilde{y}_j(T) = \frac{\sum_{t=1}^{T} I(j, s^C(t))}{T} \tag{11}$$

In the above equation, $I(a, b)$ is the indicator function, $I(a, b) = 1$ if $a = b$ and $I(a, b) = 0$ if $a \neq b$. Player Row then selects the action that maximizes the expected payoff at time T

$$i(T) = \mathrm{argmax}_k \sum_j \Pi^{\mathrm{R}}(k, j)\tilde{y}_j(T) \tag{12}$$

The behavior of Column is equivalent. We used the same convergence criteria and the same length of the simulation runs as in reinforcement learning. We checked convergence of the empirical distribution of actions, and not of actual actions, as the latter would be impossible in cyclic games (*54*). There are no parameters in fictitious play.

Replicator dynamics

We simulated the discrete version proposed by Maynard Smith [(*5*), appendix D, p. 183]

$$\begin{aligned} x_i(t+1) &= x_i(t)\frac{1+\delta\sum_j \Pi^{\mathrm{R}}(i,j)y_j(t)}{1+\delta\sum_{kj} x_k(t)\Pi^{\mathrm{R}}(k,j)y_j(t)} \\ y_j(t+1) &= y_j(t)\frac{1+\delta\sum_i \Pi^{\mathrm{C}}(j,i)x_i(t)}{1+\delta\sum_{ik} y_k(t)\Pi^{\mathrm{C}}(k,i)x_i(t)} \end{aligned} \tag{13}$$

with δ = 0.1. Here, the length of the simulation run was endogenously determined by the first component of the mixed strategy vector hitting the machine precision boundary. (Because replicator dynamics is of multiplicative nature, the components drift exponentially toward the faces of the strategy space and quickly reach the machine precision boundaries.) To verify convergence, we checked if the largest component of the mixed strategy vector of each player has been monotonically increasing over the last 20% of the time steps, and if all other components have been monotonically decreasing in the same time interval.

Experience-weighted attraction

Each player μ at time t has an attraction $Q_i^{\mu}(t)$ toward action i. The attractions update as

$$Q_i^{\mu}(t+1) = \frac{(1-\alpha)\mathcal{N}(t)Q_i^{\mu}(t) + (\delta + (1-\delta)x_i^{\mu}(t))\sum_j \Pi^{\mu}(i,j)y_j(t)}{\mathcal{N}(t+1)} \tag{14}$$

where α and δ are parameters and $\mathcal{N}(t)$ is interpreted as experience. Experience updates as $\mathcal{N}(t+1) = (1-\alpha)(1-\kappa)\mathcal{N}(t) + 1$, where κ is a parameter. Attractions map to probabilities through a logit function

$$x_i^{\mu}(t+1) = \frac{e^{\beta Q_i^{\mu}(t+1)}}{\sum_j e^{\beta Q_j^{\mu}(t+1)}} \tag{15}$$

where β is a parameter. We simulated Eqs. 14 and 15 for 500 time steps, starting with $\mathcal{N}(0) = 1$. The parameter values are α = 0.18, $\beta = \sqrt{N}$, κ = 1, and δ = 1. If in the last 100 time steps the average log variation is larger than 0.01, the simulation run was identified as nonconvergent. In formula, we checked if $1/N\sum_{i=1}^{N} 5/T\sum_{t=4/5T}^{T}(\log x_i(t))^2 > 10^{-2}$, and equivalently for Column.

EWA with noise

We replace Eq. 14 by

$$Q_i^{\mu}(t+1) = \frac{(1-\alpha)\mathcal{N}(t)Q_i^{\mu}(t) + (\delta + (1-\delta)I(i, s_{\mu}(t+1))\Pi^{\mu}(i, s_{-\mu}(t+1))}{\mathcal{N}(t+1)} \tag{16}$$

i.e., we consider online learning. The parameter values are the same as in EWA. The convergence criteria are different. We ran the dynamics

for 5000 time steps and—as in reinforcement learning—we considered only the last 20% of the time steps and only the components of the mixed strategy vectors played with average probability greater than 0.05 in this time interval. We then identified the position of the fixed point, and we classified the run as nonconvergent if play was farther than 0.02 from the fixed point in more than 10% of the time steps (i.e., in at least 100 time steps).

Level-*k* learning

Let $F^R(\cdot)$ and $F^C(\cdot)$ be the EWA updates for players Row and Column, respectively, i.e., if both players use EWA, then $x(t+1) = F^R(x(t), y(t))$ and $y(t+1) = F^C(x(t), y(t))$. (x and y without a subscript indicate the full mixed strategy vector.) Then, if Column is a level 2 learner, she updates her strategies according to $y^2(t+1) = F^C(x(t+1), y(t)) = F^C(F^R(x(t), y(t)), y(t))$. Row behaves equivalently. In the simulations, we assumed that both players are level 2 and used the same parameters and convergence criteria as in EWA.

Payoff matrices

For each payoff matrix, we randomly generated N^2 pairs of payoffs—if Row plays i and Column plays j, a pair (a, b) implies that Row receives payoff a and Column gets payoff b. We then kept the payoff matrix fixed for the rest of the simulation. Each pair was randomly sampled from a bivariate Gaussian distribution with mean 0, variance 1, and covariance Γ.

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at http://advances.sciencemag.org/cgi/content/full/5/2/eaat1328/DC1

Section S1. Details of the simulation protocol
Section S2. Best reply structure and mixed strategy NE
Section S3. Further evidence on the predictive power of the best reply structure
Section S4. Ensemble of games constrained by a potential function
Section S5. Analytical calculations on the best reply structure with uncorrelated payoffs
Fig. S1. Instances of simulation runs of the Bush-Mosteller reinforcement learning algorithm with $N = 20$.
Fig. S2. Instances of simulation runs of fictitious play with $N = 20$.
Fig. S3. Effect of negative payoff correlation on fictitious play for some values of N and Γ.
Fig. S4. Instances of simulation runs of replicator dynamics with $N = 20$.
Fig. S5. Instances of simulation runs of EWA with $N = 20$.
Fig. S6. Instances of simulation runs of EWA and EWA with noise with $N = 20$.
Fig. S7. Mixed strategy NE classified in relation to the best reply structure.
Fig. S8. Test for how well the best reply structure predicts nonconvergence with $N = 5$, instead of $N = 20$ as in the main paper.
Fig. S9. Test for how well the best reply structure predicts nonconvergence with $N = 50$, instead of $N = 20$ as in the main paper.
Fig. S10. Frequency of convergence to mixed strategy NE for $N = 20$.
Fig. S11. Correlation matrix of the co-occurrence of nonconvergence for the six learning algorithms.
Fig. S12. How the share of best reply cycles predicts convergence as perfect potential games are approached.
Fig. S13. Exhaustive account of best reply configurations with the same best reply vector.
Fig. S14. Payoff matrix with $N = 11$ that is used to illustrate the calculation of the frequency of the best reply vectors.
Fig. S15. Frequency of k-cycles, $\rho(N, k)$, as a function of the number of actions N.
References (*55*–*62*)

REFERENCES AND NOTES

1. R. B. Myerson, *Game Theory* (Harvard Univ. Press, 2013).
2. R. Axelrod, W. D. Hamilton, The evolution of cooperation. *Science* **211**, 1390–1396 (1981).
3. M. A. Nowak, D. C. Krakauer, The evolution of language. *Proc. Natl. Acad. Sci. U.S.A.* **96**, 8028–8033 (1999).
4. R. W. Rosenthal, A class of games possessing pure-strategy Nash equilibria. *Int. J. Game Theory* **2**, 65–67 (1973).
5. J. Maynard Smith, *Evolution and the Theory of Games* (Cambridge Univ. Press, 1982).
6. R. M. May, Qualitative stability in model ecosystems. *Ecology* **54**, 638–641 (1973).
7. G. P. Morriss, D. J. Evans, *Statistical Mechanics of Nonequilibrium Liquids* (ANU Press, 2013).
8. D. Monderer, L. S. Shapley, Potential games. *Games Econ. Behav.* **14**, 124–143 (1996).
9. J. H. Nachbar, "Evolutionary" selection dynamics in games: Convergence and limit properties. *Int. J. Game Theory* **19**, 59–89 (1990).
10. D. P. Foster, H. P. Young, On the nonconvergence of fictitious play in coordination games. *Games Econ. Behav.* **25**, 79–96 (1998).
11. P. Milgrom, J. Roberts, Rationalizability, learning, and equilibrium in games with strategic complementarities. *Econometrica* **58**, 1255–1277 (1990).
12. I. Arieli, H. P. Young, Stochastic learning dynamics and speed of convergence in population games. *Econometrica* **84**, 627–676 (2016).
13. D. Easley, J. Kleinberg, *Networks, Crowds, and Markets: Reasoning About a Highly Connected World* (Cambridge Univ. Press, 2010).
14. R. M. May, Will a large complex system be stable? *Nature* **238**, 413–414 (1972).
15. S. Johnson, V. Domínguez-García, L. Donetti, M. A. Muñoz, Trophic coherence determines food-web stability. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 17923–17928 (2014).
16. D. P. Foster, R. V. Vohra, Calibrated learning and correlated equilibrium. *Games Econ. Behav.* **21**, 40–55 (1997).
17. S. Hart, A. Mas-Colell, A simple adaptive procedure leading to correlated equilibrium. *Econometrica* **68**, 1127–1150 (2000).
18. S. Hart, Adaptive heuristics. *Econometrica* **73**, 1401–1430 (2005).
19. A. Blum, Y. Mansour, Learning, regret minimization, and equilibria, in *Algorithmic Game Theory*, N. Nisan, T. Roughgarden, E. Tardos, V. V. Vazirani, Eds. (Cambridge Univ. Press, 2007), pp. 79–102.
20. R. J. Aumann, Correlated equilibrium as an expression of Bayesian rationality. *Econometrica* **55**, 1–18 (1987).
21. L. E. Blume, The statistical mechanics of strategic interaction. *Games Econ. Behav.* **5**, 387–424 (1993).
22. K. Goldberg, A. Goldman, M. Newman, The probability of an equilibrium point. *J. Res. Natl. Bur. Stand.* **72**, 93–101 (1968).
23. M. Dresher, Probability of a pure equilibrium point in *n*-person games. *J. Comb. Theory* **8**, 134–145 (1970).
24. I. Y. Powers, Limiting distributions of the number of pure strategy Nash equilibria in n-person games. *Int. J. Game Theory* **19**, 277–286 (1990).
25. M. Opper, S. Diederich, Phase transition and 1/f noise in a game dynamical model. *Phys. Rev. Lett.* **69**, 1616–1619 (1992).
26. J. Berg, A. Engel, Matrix games, mixed strategies, and statistical mechanics. *Phys. Rev. Lett.* **81**, 4999–5002 (1998).
27. J. Berg, Statistical mechanics of random two-player games. *Phys. Rev. E* **61**, 2327–2339 (2000).
28. J. E. Cohen, Cooperation and self-interest: Pareto-inefficiency of Nash equilibria in finite random games. *Proc. Natl. Acad. Sci. U.S.A.* **95**, 9724–9731 (1998).
29. T. Galla, J. D. Farmer, Complex dynamics in learning complicated games. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 1232–1236 (2013).
30. S. A. Kauffman, Metabolic stability and epigenesis in randomly constructed genetic nets. *J. Theor. Biol.* **22**, 437–467 (1969).
31. B. Skyrms, Chaos in game dynamics. *J. Logic Lang. Inf.* **1**, 111–130 (1992).
32. L. E. Blume, D. Easley, Learning to be rational. *J. Econ. Theory* **26**, 340–351 (1982).
33. M. Boldrin, L. Montrucchio, On the indeterminacy of capital accumulation paths. *J. Econ. Theory* **40**, 26–39 (1986).
34. C. Hommes, G. Sorger, Consistent expectations equilibria. *Macroecon. Dyn.* **2**, 287–321 (1998).
35. G. Gigerenzer, P. M. Todd, *Simple Heuristics That Make Us Smart* (Oxford Univ. Press, 1999).
36. I. Erev, A. E. Roth, Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. *Am. Econ. Rev.* **88**, 848–881 (1998).
37. R. R. Bush, F. Mosteller, *Stochastic Models for Learning* (John Wiley & Sons Inc., 1955).
38. J. Robinson, An iterative method of solving a game. *Ann. Math.* 296–301 (1951).
39. G. W. Brown, *Activity Analysis of Production and Allocation*, T. Koopmans, Ed. (Wiley, 1951), pp. 374–376.
40. Y.-W. Cheung, D. Friedman, Individual learning in normal form games: Some laboratory results. *Games Econ. Behav.* **19**, 46–76 (1997).
41. D. O. Stahl, On the instability of mixed-strategy Nash equilibria. *J. Econ. Behav. Organ.* **9**, 59–69 (1988).
42. J. Hofbauer, K. Sigmund, *Evolutionary Games and Population Dynamics* (Cambridge Univ. Press, 1998).
43. T. Börgers, R. Sarin, Learning through reinforcement and replicator dynamics. *J. Econ. Theory* **77**, 1–14 (1997).
44. C. Camerer, T.-H. Ho, Experience-weighted attraction learning in normal form games. *Econometrica* **67**, 827–874 (1999).

45. M. Pangallo, J. B. Sanders, T. Galla, J. D. Farmer, A taxonomy of learning dynamics in 2 x 2 games. arXiv:1701.09043 [q-fin.EC] (31 January 2017).
46. R. Nagel, Unraveling in guessing games: An experimental study. *Am. Econ. Rev.* **85**, 1313–1326 (1995).
47. R. Selten, *Game Equilibrium Models I* (Springer-Verlag, 1991), pp. 98–154.
48. V. P. Crawford, M. A. Costa-Gomes, N. Iriberri, Structural models of nonequilibrium strategic thinking: Theory, evidence, and applications. *J. Econ. Lit.* **51**, 5–62 (2013).
49. F.-F. Tang, Anticipatory learning in two-person games: Some experimental results. *J. Econ. Behav. Organ.* **44**, 221–232 (2001).
50. K. Miyazawa, "On the convergence of the learning process in a 2 × 2 non-zero-sum two person game" (Technical Report Research Memorandum No. 33, Econometric Research Program, Princeton University, 1961).
51. L. S. Shapley, Some topics in two-person games. *Adv. Game Theory* **52**, 1–29 (1964).
52. H. Gintis, *Game Theory Evolving: A Problem-Centered Introduction to Modeling Strategic Behavior* (Princeton Univ. Press, 2000).
53. J. B. Sanders, J. D. Farmer, T. Galla, The prevalence of chaotic dynamics in games with many players. *Sci. Rep.* **8**, 4902 (2018).
54. D. Fudenberg, D. K. Levine, *The Theory of Learning in Games* (MIT Press, 1998) vol. 2.
55. V. P. Crawford, Learning the optimal strategy in a zero-sum game. *Econometrica* 885–891 (1974).
56. J. Conlisk, Adaptation in games: Two solutions to the Crawford puzzle. *J. Econ. Behav. Organ.* **22**, 25–50 (1993).
57. R. Bloomfield, Learning a mixed strategy equilibrium in the laboratory. *J. Econ. Behav. Organ.* **25**, 411–436 (1994).
58. M. W. Macy, A. Flache, Learning dynamics in social dilemmas. *Proc. Natl. Acad. Sci. U.S.A.* **99** (suppl. 3), 7229–7236 (2002).
59. Y. Sato, E. Akiyama, J. P. Crutchfield, Stability and diversity in collective adaptation. *Phys. D Nonlin. Phenom.* **210**, 21–57 (2005).
60. D. Lecutier, "Stochastic dynamics of game learning," thesis, University of Manchester (2013).
61. T. Evans, "k-level reasoning: A dynamic model of game learning," thesis, University of Manchester (2013).
62. N. Nisan, T. Roughgarden, E. Tardos, V. V. Vazirani, *Algorithmic Game Theory* (Cambridge Univ. Press, 2007) vol. 1.

Acknowledgments

Funding: This work was supported by INET and EPSRC award number 1657725. **Author contributions:** M.P. conceived the research. M.P., T.H., and J.D.F. designed the analyses. M.P. and T.H. conducted the analyses. All authors wrote the manuscript. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. The code to reproduce the quantitative results in the paper is available on Zenodo (https://dx.doi.org/10.5281/zenodo.1419542). Additional data related to this paper may be requested from the authors.

Submitted 17 March 2018
Accepted 8 January 2019
Published 20 February 2019
10.1126/sciadv.aat1328

Citation: M. Pangallo, T. Heinrich, J. Doyne Farmer, Best reply structure and equilibrium convergence in generic games. *Sci. Adv.* **5**, eaat1328 (2019).