THE EVOLUTIONARY ORIGIN OF SOCIAL REWARDS

1. Model

1.1. **Game, payoffs, and utilities.** We consider a general N-player game where each player i has an action set $\mathcal{A}_i$ from which it can choose an action $a_i \in \mathcal{A}_i$. We think of these N players as belonging to a larger population, where a (random) matching has occurred between population members that generated many groups of N players (see below the evolutionary setting). The action space $\mathcal{A}_i$ can be a finite set of n discrete actions or a continuous set, such as a subset of $\mathbb{R}$. The environment as well as the actions of the other players affect the material payoff of every individual. There is a set Θ of environmental states, with typical element $\theta \in \Theta$ (the environment can also be continuous or discrete). The material payoff of a player is given by the function $\pi_i : \prod_{j\in N} \mathcal{A}_j \times \Theta \to \mathbb{R}$, which ultimately affects the fitness of i (see below for details). To simplify notation, we will also write $\mathcal{A} = \prod_{j\in N} \mathcal{A}_j$ for the set of all possible action profiles.

The players have subjective preference or utility functions, $u_i : \prod_{j\in N} \mathcal{A}_j \times \Theta \to \mathbb{R}$, that may differ from the objective material payoff π_i. The utility function u_i is genetically determined and is the evolving trait: we are interested in the function(s) u_i that is(are) favored by natural selection. In the forthcoming analysis of the model, it may be easier to think of the game as having a single-valued "outcome", o that is a function of the action profile and the environment, i.e. $o(\mathbf{a}, \theta) \in O$, where the outcome space, O, is a subset of $\mathbb{R}$. The utility function may then be defined directly over outcomes $u_i : O \to \mathbb{R}$. For example, in a public-goods game, the outcome may be defined as the group productivity (e.g. the sum of all players' contributions multiplied by a synergy factor). The reason we make these additional definitions about game outcomes is that it may be easier to find general evolutionarily stable utility functions that are univariate (when the outcome is single-valued) than multivariate (when the outcome is the combination of the action profile and the state of the environment), and we could use the techniques of Kirkpatrick & Heckman (1989) to perform such an analysis.

The game is repeated at discrete times $t = 1, 2, \ldots$ and the players choose an action profile $\mathbf{a}_t$. The environment fluctuates independently from the players' actions and takes value θ_t at time t. The probability that state θ occurs at time t is written $\rho(\theta)$, and is thus independent of time (i.i.d. environment). The players cannot observe the state of the environment prior to choosing their actions. The material payoff to player i when the action profile is $\mathbf{a}_t$ and the environment is in state θ_t is given by $\pi_i(\mathbf{a}_t, \theta_t)$, which we may also write $\pi_i(a_{i,t}, \mathbf{a}_{-i,t}, \theta_t)$, where $\mathbf{a}_{-i,t} \in \prod_{\substack{j\in N \\ j\neq i}} \mathcal{A}_j$ is the action profile of all players except player i. Each individual observes privately the utility $u_i(\mathbf{a}_t, \theta_t)$ which results from the game outcome $o_t = (\mathbf{a}_t, \theta_t)$ at time t.

1.2. **Learning.** The learning model is taken from Dridi & Lehmann (2014), where players use the material payoffs of the game to update motivations about actions. Our learning model will not be very different from this previous work because all we have to change is that individuals update their choice probabilities using the subjective utility of a game outcome, rather than the objective material payoff. Hence, the difference between this previous paper and our formulation of learning dynamics is that in Dridi & Lehmann (2014), the games considered are symmetric,

Date: March 4, 2016.

while here the utility function of each player is different, which translates into an asymmetric game defined by the family of utility functions $(u_i)_{i\in N}$.

While the general model of Dridi & Lehmann (2014) allows for both trial-and-error learning and belief-based learning, we will first only consider trial-and-error learning. At every time t of the repeated game defined above, each individual i holds in memory action values $V_{i,t}(a_i)$ for all actions $a_i \in \mathcal{A}_i$. The learning rule of individual i is to update action values according to

$$V_{i,t+1}(a_i) = (1-\gamma_t)V_{i,t}(a_i) + \gamma_t \mathbb{1}(a_i, a_{i,t})u_i(a_i, \mathbf{a}_{-i,t}, \theta_t), \tag{1}$$

where $\gamma_t \in (0,1)$ is a decreasing learning rate in the sense of stochastic approximation theory. This decreasing learning rate allows us to approximate the above stochastic difference equation with a deterministic mean-field differential equation that asymptotically tracks the original stochastic dynamics. The expression $\mathbb{1}(a_i, a_{i,t})$ is an indicator variable that equals 1 if $a_i = a_{i,t}$, and 0 otherwise.

An alternative learning rule, whose mean field is an "unperturbed" version of the mean field of eq. 1 (which is thus easier to analyze; see Dridi & Lehmann, 2015, for details), can be written as

$$V_{i,t+1}(a_i) = V_{i,t}(a_i) + \gamma_t \mathbb{1}(a_i, a_{i,t})u_i(a_i, \mathbf{a}_{-i,t}, \theta_t). \tag{2}$$

While eq. 2 is less used in the literature, it corresponds more to a conscious updating of action values than eq. 1 because in eq. 2 an action that is not played at time t keeps the same action value at time $t+1$. Under eq. 1 action values tend to decrease for non-played actions.

We generally call the right-hand side of eqs. 1–2 the learning rule, written $\ell_i(V_i, h_i)$, of individual i. The learning rule takes the previous vector of action values V_i and a new information h_i (in our present model, $h_i = u_i(\cdot, \mathbf{a}_{-i,t}, \theta_t)$, but one could think of more general updating rules) and outputs a new vector of action values.

Player i chooses an action $a_{i,t}$ at time t with a probability that depends on its action values $V_{i,t} = \{V_{i,t}(a)\}_{a\in\mathcal{A}}$. One possibility is to assume a perturbed maximization scheme which gives rise to the logit-choice function,

$$p_{i,t}(a_i) = \frac{\exp[\xi V_{i,t}(a_i)]}{\sum_{b_i\in\mathcal{A}_i}\exp[\xi V_{i,t}(b_i)]}, \tag{3}$$

where ξ is the exploration parameter or noise level in choosing actions. We write $p_{i,t}$ without the action argument to denote the whole vector of action probabilities of player i.

2. FECUNDITY

We define the total payoff of individual i as the average material payoff obtained at equilibrium of the learning process, i.e.

$$f_i = f(u_i) = \int_{\mathcal{A}} \sum_{\theta\in\Theta} \hat{\mathbf{p}}(\mathbf{a})\rho(\theta)\pi_i(\mathbf{a},\theta)\,\mathrm{d}\mathbf{a}, \tag{4}$$

where $\hat{\mathbf{p}}(\mathbf{a}) = \prod_{j\in N}\hat{p}_j(a_j)$ is the equilibrium probability of action profile $\mathbf{a} = (a_1, \dots, a_N)$. This is the product of individuals' equilibrium action probabilities $\hat{p}_j(a_j)$.

Importantly, while the utility function does not appear on the rhs of eq. 4, we still defined it as $f(u_i)$ because the equilibrium choice probabilities of a player, $\hat{p}_i(a_i)$, will generally depend on the utility function of player i.

3. Preference evolution and the set of possible utility functions

Consider a very large population and assume that groups of N players are randomly formed at every generation to play the above repeated game. An individual's genotype corresponds to its utility function u_i, which he transmits to its offspring. The number of offspring he produces depends on its material payoffs obtained during the game as defined in eq. 4.

In the most general case, the set in which evolution occurs is the set of all possible utility functions $u : \prod_{j \in N} \mathcal{A}_j \times \Theta \to \mathbb{R}$. Let U denote such a set.

4. Analysis of simple cases

4.1. Symmetric two-player games.

It can be shown that a trial-and-error learner's equilibrium action is such that $u_i(\hat{a}_i, \cdot) \geq 0$ (Dridi & Lehmann, 2015). In other words, as long as the utility obtained when playing action $\hat{a}_i$ is positive, then $\hat{a}_i$ is an equilibrium action (we wrote here $\hat{a}_i$ for the action that has probability one under the equilibrium of the learning dynamics, i.e. $\hat{p}_i(\hat{a}_i) = 1$). As a consequence, there are $2^4 = 16$ possible types of utility functions that are relevant here because, for each of the 4 game outcomes (a_i, a_{-i}), the individual may have a positive or negative utility for this outcome: $u_i(a_i, a_{-i}) \geq 0$ or $u_i(a_i, a_{-i}) < 0$.

This result is nice because this means that we would not need to make any restrictive assumption on the set U of allowable utility functions. Statements regarding evolutionary stability will be here in terms of families of utility functions. Of course, the cost of this approach is that results won't be explicit about the exact functional form that the ESS utility function will take.

4.1.1. *Locally stable utility functions.*

The set U consists of all possible 2×2 real matrices, which can be identified with the set $\mathbb{R}^{2\times 2} = \mathbb{R}^4$. The genotype of an individual can thus also be seen as the vector $\mathbf{u} = (u_{11}, u_{12}, u_{21}, u_{22})$, where u_{ij} denotes the utility to the focal player when he chooses action i and his opponent chooses action j.

With such definitions, one can use standard approaches in evolutionary biology, such as adaptive dynamics, in order to find the ESS utility matrix, and the analysis can be complemented to look at convergence stability. Under this approach one considers a resident population with utility $\mathbf{u}$ and asks whether a mutant with utility $\mathbf{u} + \boldsymbol{\delta}$ can invade or not. One advantage of this approach in our learning context is that the "utility game" between the resident and the mutant is quasi-symmetric, which implies that the learning dynamics determining behavior is a family of dynamics for quasi-symmetric games. This is a substantial simplification of the analysis. Save the bifurcations, the quasi-symmetric game between $\mathbf{u}$ and $\mathbf{u} + \boldsymbol{\delta}$ is very close to the symmetric game defined by $\mathbf{u}$, because the vector field determining the learning dynamics is continuous in $\mathbf{u}$.

A candidate ESS is such that

$$\left.\frac{\partial f}{\partial \mathbf{u}}\right|_{\mathbf{u}=\mathbf{u}^*} = 0 \tag{5}$$

4.2. Subset of strategies in the Prisoner's Dilemma.

Of the 7 strategies considered in Fig. 1, the Realistic strategy (and the equivalent Avoid Sucker's) is the "winning" strategy (in a sense to be made precise later). Moreover it seems that among the set of Realistic strategies, those that put higher values on positive rewards would be favored, because they lead to faster learning (TO DO: check this). This sheds light on the maintenance of addiction to positive rewards in human populations. Overestimation of rewards (i.e. give to a reward more value than its real fitness value) seems to be evolutionarily stable.

TABLE 1. Classification of behavioral outcomes in the discrete action model amongst the 4 strategies considered.

Interaction	Always stable equilibria	Sometimes stable equilibria	Stability condition
R vs. R	$(1, 1)$ and $(0, 0)$		
R vs. O	$(0, 1)$ and $(1, 1)$		
R vs. M	$(1, 1)$ and $(0, \frac{v_{22}}{v_{12}+v_{22}})$		
R vs. S	$(0, 0)$		
O vs. O	$(1, 1)$		
O vs. M	$(1, 0)$ and $(1, 1)$		
O vs. S	$(1, 0)$		
M vs. M	$(1, 1)$	$(\frac{u_{22}}{u_{12}+u_{22}}, 0)$ and $(0, \frac{v_{22}}{v_{12}+v_{22}})$	$\lvert u_{12}\rvert < \lvert u_{21}\rvert$, $\lvert v_{12}\rvert < \lvert v_{21}\rvert$,
M vs. S	$(\frac{u_{22}}{u_{12}+u_{22}}, 0)$		
S vs. S	$(0, 0)$		

The strategy name “Manipulator” stems from the fact that an individual using this strategy will drive an indifferent opponent (i.e. that has positive utility for all four game outcomes) to cooperate, while the Manipulator against the indifferent opponent may converge to cooperation or defection depending on the stochastic events occurring during the game interaction.

4.3. **Analysis of the behavioral interactions between the strategies.** In this section we calculate the payoffs and fitnesses for all possible behavioral interactions between the 4 strategies Realistic, Other-regard, Selfish, and Manipulator. From these calculations, we derive invasion conditions within this set of strategies.

Some assumptions need to be made in order to compute analytical expressions for long-term payoffs. First, when the behavioral dynamics admit several stable equilibria, typically the stochastic dynamics may converge to any of these stable equilibria. It is very difficult to know which particular equilibrium will be reached by the stochastic process, because the trajectory may go out from the basin of attraction of a stable equilibrium under the influence of large stochastic shocks (these large shocks are more likely to occur at the beginning of the behavioral interaction). For these reasons, we assume that the initial preferences of players are unbiased, such that the process starts in the center of the state space ($p_1 = p_2 = 1/2$). Also, we assume that the process may reach each equilibrium with equal probability, in other words the distribution of stable equilibria is uniform. Such an assumption is valid for values of the exploration rate around $\xi \approx 10$ (Dridi & Lehmann, 2015).

TABLE 2. Generic total payoff in the discrete action model amongst the 4 strategies considered.

Interaction	Generic fitness
R vs. R	$(\frac{1}{2}(\mathcal{R}+\mathcal{P}), \frac{1}{2}(\mathcal{R}+\mathcal{P}))$
R vs. O	$(\frac{1}{2}(\mathcal{R}+\mathcal{T}), \frac{1}{2}(\mathcal{R}+\mathcal{S}))$
R vs. M	$\left(\frac{1}{2}\left(\mathcal{R}+\mathcal{T}\left(\frac{v_{22}}{v_{12}+v_{22}}\right)+\mathcal{P}\left(1-\frac{v_{22}}{v_{12}+v_{22}}\right)\right), \frac{1}{2}\left(\mathcal{R}+\mathcal{S}\left(\frac{v_{22}}{v_{12}+v_{22}}\right)+\mathcal{P}\left(1-\frac{v_{22}}{v_{12}+v_{22}}\right)\right)\right)$
R vs. S	$(\mathcal{P}, \mathcal{P})$
O vs. O	$(\mathcal{R}, \mathcal{R})$
O vs. M	$(\frac{1}{2}(\mathcal{R}+\mathcal{S}), \frac{1}{2}(\mathcal{R}+\mathcal{T}),)$
O vs. S	$(\mathcal{S}, \mathcal{T})$
M vs. M	$(\mathcal{R}, \mathcal{R})$ or ...
M vs. S	$\left(\mathcal{S}\left(\frac{u_{22}}{u_{12}+u_{22}}\right)+\mathcal{P}\left(1-\frac{u_{22}}{u_{12}+u_{22}}\right), \mathcal{T}\left(\frac{u_{22}}{u_{12}+u_{22}}\right)+\mathcal{P}\left(1-\frac{u_{22}}{u_{12}+u_{22}}\right)\right)$
S vs. S	$(\mathcal{P}, \mathcal{P})$

TABLE 3. Total payoff in the 3 different types of games in the discrete action model amongst the 4 strategies considered.

Interaction	Prisoner's Dilemma	Stag Hunt	Snowdrift
R vs. R	$(\frac{b-c}{2}, \frac{b-c}{2})$	$(\frac{(k+1)(b-c)}{2k}, \frac{(k+1)(b-c)}{2k})$	$(\frac{2b-c}{4}, \frac{2b-c}{4})$
R vs. O	$(b-\frac{c}{2}, \frac{b-c}{2})$	$(\frac{(k+1)(b-c)}{2k}, \frac{b-c}{2})$	$(b-\frac{c}{4}, b-\frac{3c}{4})$
R vs. M	$(\frac{(b-c)v_{12}+(2b-c)v_{22}}{2(v_{12}+v_{22})}, \frac{(b-c)v_{12}+(b-2c)v_{22}}{2(v_{12}+v_{22})})$	$(\frac{(b-c)k+(b-c)}{2k}, \frac{(b-c+(b-c)k)v_{12}+(b-2c)kv_{22}}{2k(v_{12}+v_{22})})$	$(\frac{(2b-c)v_{12}+(4b-c)v_{22}}{4(v_{12}+v_{22})}, \frac{(2b-c)v_{12}+(4b-3c)v_{22}}{4(v_{12}+v_{22})})$
R vs. S	$(0,0)$	$(\frac{b-c}{k}, \frac{b-c}{k})$	$(0,0)$
O vs. O	$(b-c, b-c)$	$(b-c, b-c)$	$(b-\frac{c}{2}, b-\frac{c}{2})$
O vs. M	$(\frac{b}{2}-c, b-\frac{c}{2})$	$(\frac{b}{2}-c, \frac{(b-c)k+b-c}{2k})$	$(b-\frac{3c}{4}, b-\frac{c}{4})$
O vs. S	$(-c, b)$	$(-c, \frac{b-c}{k})$	$(b-c, b)$
M vs. M	$(b-c, b-c)$	$(b-c, b-c)$	$(b-\frac{c}{2}, b-\frac{c}{2})$
M vs. S	$(\frac{-cu_{22}}{u_{12}+u_{22}}, \frac{bu_{22}}{u_{12}+u_{22}})$	$(\frac{(b-c)u_{12}-cku_{22}}{k(u_{12}+u_{22})}, \frac{b-c}{k})$	$(\frac{(b-c)u_{22}}{u_{12}+u_{22}}, \frac{bu_{22}}{u_{12}+u_{22}})$
S vs. S	$(0,0)$	$(\frac{b-c}{k}, \frac{b-c}{k})$	$(0,0)$

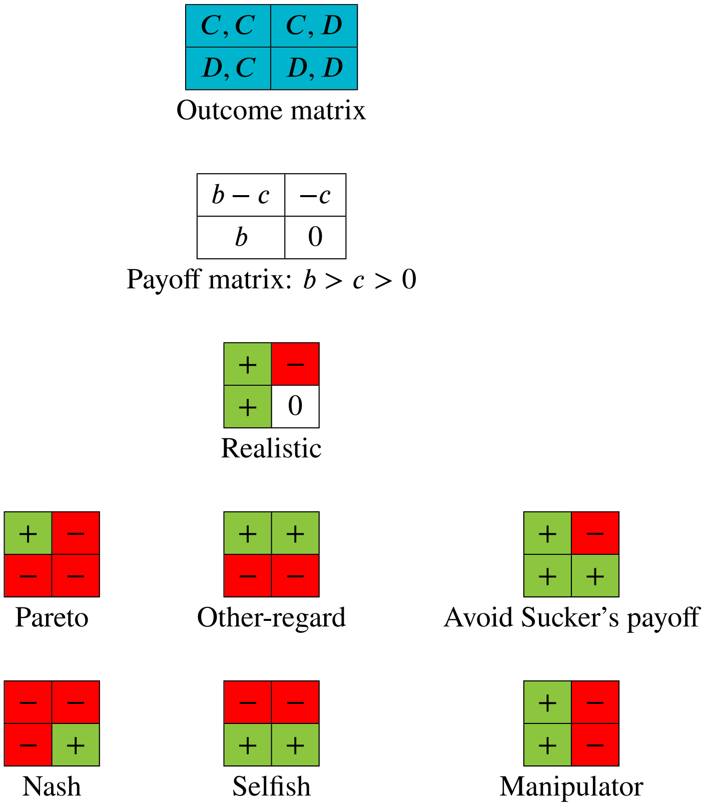

FIGURE 1. The 7 strategies considered in the discrete action model. A strategy is defined by the outcomes to which it associates a positive or negative utility in the corresponding outcome matrix (top). Cooperation is denoted by C and defection by D. In the outcome matrix, the first letter refers to the action of the focal player (row) and the second letter to the action of its opponent (column). For example, the strategy Realistic associates a positive utility to the outcomes that yield positive material payoffs, and negative utility to outcomes yielding negative material payoffs. The strategy Pareto associates a positive utility to the outcome where both players cooperate (C, C) and has a negative utility for all other outcomes.

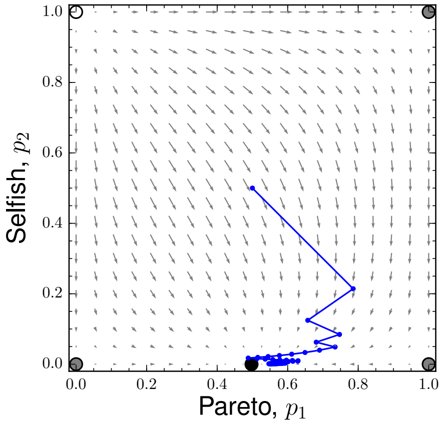

FIGURE 2. Vector field (gray arrows) and stochastic trajectory (blue line) for the interaction between "Pareto" and "Selfish". On the x-axis is represented the probability that Pareto cooperates (p_1), while on the y-axis, this is the probability that Selfish cooperates (p_2). The stochastic trajectory is started from the center of the state space $(p_1, p_2) = (\frac{1}{2}, \frac{1}{2})$ and dots on it represent interaction rounds between the players. Circles represent equilibria: a white-filled circle is a source (both associated eigenvalues are positive); a gray-filled circle is a saddle (one positive and one negative associated eigenvalue); a black circle is a sink (both associated eigenvalues are negative).

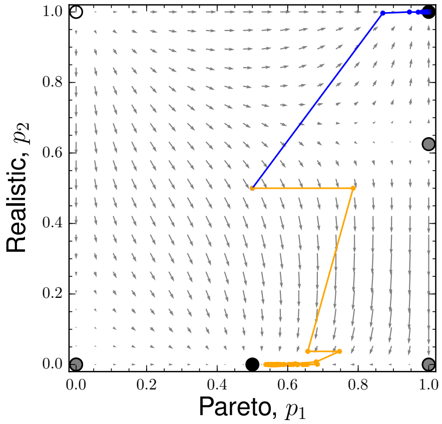

FIGURE 3. Same as Fig. 2 but for the interaction between Pareto and Realistic. Here the deterministic mean field equation admits two locally stable equilibria. The two stochastic trajectories of different color correspond to simulation runs that respectively converge to one of the locally stable equilibria.

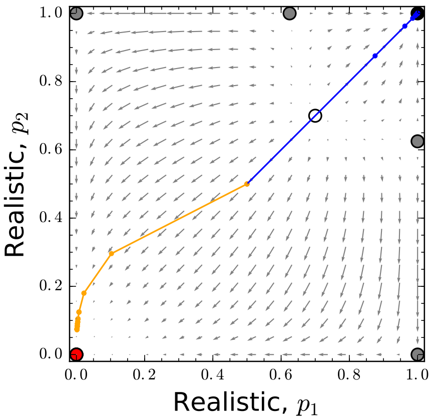

FIGURE 4. Same as Fig. 2 but for the interaction between Realistic and Realistic. The red-filled dot denotes an equilibrium with 0 eigenvalues.

5. Continuous action space

5.1. **Normally distributed actions.** Here, we treat the case where the action space is one-dimensional continuous, e.g., when $\mathcal{A}_i \subseteq \mathbb{R}$ for all i. This allows to capture investment games such as a continuous public goods game. We adapt a model of Beigy & Meibody (2006), who study learning of a single decision-maker who faces a stochastic stationary environment. In their setting, the decision-maker tries to minimize the expectation of a given cost function where at each discrete time step he can exert an action $a \in \mathbb{R}$. In order to adapt their model to multi-player games, only a few tweaks need to be made.

To start with, we consider that at each time step t, every player is characterized by a mean action $\mu_{i,t}$, and a standard deviation $\sigma_{i,t}$. The action $a_{i,t}$ chosen by i at time t is then drawn from a normal distribution with mean $\mu_{i,t}$ and standard deviation $\sigma_{i,t}$, i.e., $a_{i,t} \sim \mathcal{N}(\mu_{i,t}, \sigma_{i,t})$. A possible interpretation of this model of action choice is that $\mu_{i,t}$ represents the preferred action by player i, which he would like to implement but there is a stochastic perturbation during action choice, such that $a_{i,t} = \mu_{i,t} + \varepsilon_{i,t}$, where $\varepsilon_{i,t} \sim \mathcal{N}(0, \sigma_{i,t})$ is independent across players and time steps. With this interpretation, $\varepsilon_{i,t}$ can be seen as an exogenous noise during action implementation, or can also be seen as explicit exploration of the player around its preferred action (note that this model of action perturbation is different from the above model of payoff perturbation that generates the logit choice rule, eq. 3).

Since a normal distribution is fully specified by its mean and standard deviation, in order to know the action distribution of a player at time $t + 1$, we only need to know how $\mu_{i,t+1}$ and $\sigma_{i,t+1}$ are updated given the previous values at time t and the outcome of the game. We will use the following updating rule for the mean action,

$$\mu_{i,t+1} = \mu_{i,t} + \gamma_t u_i(a_{i,t}, \mathbf{a}_{-i,t}, \theta_t) \left(\frac{a_{i,t} - \mu_{i,t}}{\sigma_{i,t}} \right), \tag{6}$$

which essentially entails that the individual wants his new mean action $\mu_{i,t+1}$ to come closer to the chosen action $a_{i,t}$ if the obtained utility $u_i(a_{i,t}, \mathbf{a}_{-i,t}, \theta_t)$ is positive, while he wants his mean action to move away from $a_{i,t}$ if the obtained utility is negative; the magnitude of such movement is proportional to the actual magnitude of the obtained utility. Consequently, this is one of the simplest way of capturing in a continuous action setting the biological notion of approach towards rewards and avoidance of punishments. The standard deviation $\sigma_{i,t}$ appears in the denominator of the reinforcement in eq. 6 so that the individual accounts for its own exploration strategy when updating action value: if $\sigma_{i,t}$ is large, then there is a big chance that $a_{i,t}$ will be far from the mean, so the utility obtained by the individual should not be interpreted as reflecting a wrong mean action $\mu_{i,t}$ (the utility is mainly a reflection of the great variance in action choice), hence the update should be minor. Finally, the learning rate γ_t must obey the assumptions of stochastic approximation theory, just as in the discrete-action model (eq. 2) in order to ensure convergence of the learning process.

In the case where $\sigma_{i,t} = \sigma_i$ is a constant, independent of time, then σ_i acts as an additional learning rate, that scales the speed at which the mean-field deterministic dynamics converge to an equilibrium: larger variance should lead to faster convergence (see the analysis below for details).

A technical assumption needed to make the analysis tractable is that $u_i(a_{i,t}, \mathbf{a}_{-i,t}, \theta_t)$ is twice continuously differentiable in the actions of all the players (note that this may create some problems when defining games with bounded action space, where typically the payoff function is not differentiable on the boundaries).

The vector $\boldsymbol{\mu}_t = (\mu_{1,t}, \dots, \mu_{N,t})$ collects the mean action of all players at time t.

5.2. **Stochastic approximation.** Let $R_{i,t+1} = u_i(a_{i,t}, \mathbf{a}_{-i,t}, \theta_t)\left(\frac{a_{i,t}-\mu_{i,t}}{\sigma_{i,t}}\right)$ be the reinforcement to the mean action of player i, and let $\mathbf{R}_{t+1} = (R_{i,t+1})_{i\in N}$ denote the vector collecting the reinforcements of all the players at time $t+1$. Stochastic approximation theory tells us that $\mathbb{E}[\mathbf{R}_{t+1}|\boldsymbol{\mu}_t]$ is a vector field that defines a differential equation for $\boldsymbol{\mu}$ whose asymptotic path is followed by the original stochastic process $\{\boldsymbol{\mu}_t\}_{t\geq 0}$. To obtain an explicit expression for $\mathbb{E}[\mathbf{R}_{t+1}|\boldsymbol{\mu}_t]$, we use a second-order Taylor series expansion of $u_i(a_{i,t}, \mathbf{a}_{-i,t}, \theta_t)$ around $\mu_{i,t}$, which gives

$$\mathbb{E}[R_{i,t+1}|\boldsymbol{\mu}_t] = \mathbb{E}\Big[\Big(u_i(\mu_{i,t}, \boldsymbol{\mu}_{-i,t}, \theta_t) + \nabla u_i(\mu_{i,t}, \boldsymbol{\mu}_{-i,t}, \theta_t)^{\mathrm{T}}(\mathbf{a}_t - \boldsymbol{\mu}_t) \\ + (\mathbf{a}_t - \boldsymbol{\mu}_t)^{\mathrm{T}}\mathbf{H}u_i(\mu_{i,t}, \boldsymbol{\mu}_{-i,t}, \theta_t)(\mathbf{a}_t - \boldsymbol{\mu}_t)\Big)\frac{(a_{i,t}-\mu_{i,t})}{\sigma_i}\Big|\boldsymbol{\mu}_t\Big], \quad (7)$$

where $\nabla u_i(\mu_{i,t}, \boldsymbol{\mu}_{-i,t}, \theta_t) = \left(\frac{\partial u_i}{\partial \mu_{1,t}}, \ldots, \frac{\partial u_i}{\partial \mu_{N,t}}\right)$ is the gradient vector and $\mathbf{H}u_i(\mu_{i,t}, \boldsymbol{\mu}_{-i,t}, \theta_t)$ is the Hessian matrix of u_i, with typical element $[\mathbf{H}u_i]_{kj} = \frac{\partial^2 u_i}{\partial \mu_{k,t}\partial \mu_{j,t}}$. The vector $\mathbf{a}_t - \boldsymbol{\mu}_t$ has typical element $a_{j,t} - \mu_{j,t}$. Taking the non-random terms out of the expectation, the above equation can be rewritten as

$$\mathbb{E}[R_{i,t+1}|\boldsymbol{\mu}_t] = \frac{1}{\sigma_i}u_i(\mu_{i,t}, \boldsymbol{\mu}_{-i,t}, \theta_t)\mathbb{E}\Big[(a_{i,t}-\mu_{i,t})\Big|\boldsymbol{\mu}_t\Big] + \nabla u_i(\mu_{i,t}, \boldsymbol{\mu}_{-i,t}, \theta_t)\mathbb{E}\Big[(a_{i,t}-\mu_{i,t})(\mathbf{a}_t - \boldsymbol{\mu}_t)\Big|\boldsymbol{\mu}_t\Big] \\ + \mathbb{E}\Big[(a_{i,t}-\mu_{i,t})(\mathbf{a}_t - \boldsymbol{\mu}_t)^{\mathrm{T}}\mathbf{H}u_i(\mu_{i,t}, \boldsymbol{\mu}_{-i,t}, \theta_t)(\mathbf{a}_t - \boldsymbol{\mu}_t)\Big|\boldsymbol{\mu}_t\Big]. \quad (8)$$

In order to find expressions for the various expectations above it will be useful to keep in mind that for each player i, the random variable $a_{i,t} - \mu_{i,t} = \varepsilon_{i,t}$ is normally distributed with zero mean and variance σ_i^2. In particular, this implies that $\mathbb{E}\Big[(a_{i,t}-\mu_{i,t})\Big|\boldsymbol{\mu}_t\Big] = 0$ for all i, so that $\mathbb{E}\Big[(\mathbf{a}_t - \boldsymbol{\mu}_t)\Big|\boldsymbol{\mu}_t\Big] = (0, 0, \ldots, 0)$. Also

$$\begin{aligned}\mathbb{E}\Big[(a_{i,t}-\mu_{i,t})(\mathbf{a}_t - \boldsymbol{\mu}_t)\Big|\boldsymbol{\mu}_t\Big] &= \Big(\mathrm{Cov}[a_{1,t}, a_{i,t}|\boldsymbol{\mu}_t], \ldots, \mathrm{Var}[a_{i,t}|\boldsymbol{\mu}_t], \ldots, \mathrm{Cov}[a_{N,t}, a_{i,t}|\boldsymbol{\mu}_t]\Big) \\ &= \Big(0, \ldots, \sigma_i^2, \ldots, 0\Big), \end{aligned} \quad (9)$$

because action choice is independent across players. An interesting note here on social learning is that if players copy each other then $\mathrm{Cov}[a_{j,t}, a_{i,t}|\boldsymbol{\mu}_t]$ of players i and j may be different than 0, depending on whether they can copy one another. The term involving the Hessian matrix has the form

$$\mathbb{E}\Big[(a_{i,t}-\mu_{i,t})(\mathbf{a}_t - \boldsymbol{\mu}_t)^{\mathrm{T}}\mathbf{H}u_i(\mu_{i,t}, \boldsymbol{\mu}_{-i,t}, \theta_t)(\mathbf{a}_t - \boldsymbol{\mu}_t)\Big|\boldsymbol{\mu}_t\Big] = \\ \sum_{j\in N}\sum_{k\in N}\frac{\partial^2 u_i}{\partial \mu_{k,t}\partial \mu_{j,t}}\mathbb{E}\Big[(a_{i,t}-\mu_{i,t})(a_{j,t}-\mu_{j,t})(a_{k,t}-\mu_{k,t})\Big|\boldsymbol{\mu}_t\Big], \quad (10)$$

where the triplet covariance $\mathbb{E}\Big[(a_{i,t}-\mu_{i,t})(a_{j,t}-\mu_{j,t})(a_{k,t}-\mu_{k,t})\Big|\boldsymbol{\mu}_t\Big]$ is zero again because action choice is independent across players whenever at least one of $j \neq i$ or $k \neq i$ or $k \neq j$ is true. When $j = i$ and $k = i$, this triplet equals $\mathbb{E}\Big[(a_{i,t}-\mu_{i,t})^3\Big|\boldsymbol{\mu}_t\Big] = 0$. Thus we have

$$\mathbb{E}\Big[(a_{i,t}-\mu_{i,t})(\mathbf{a}_t - \boldsymbol{\mu}_t)^{\mathrm{T}}\mathbf{H}u_i(\mu_{i,t}, \boldsymbol{\mu}_{-i,t}, \theta_t)(\mathbf{a}_t - \boldsymbol{\mu}_t)\Big|\boldsymbol{\mu}_t\Big] = 0. \quad (11)$$

Plugging back the expressions of these expectations in eq. 8, the expected reinforcement of player i takes the simple form

$$\mathbb{E}[R_{i,t+1}|\boldsymbol{\mu}_t] = \sigma_i\frac{\partial u_i(\mu_{i,t}, \boldsymbol{\mu}_{-i,t}, \theta_t)}{\partial \mu_{i,t}}. \quad (12)$$

Thus, the differential equation

$$\frac{\mathrm{d}\mu_i}{\mathrm{d}t} \equiv \dot{\mu}_i = \sigma_i\frac{\partial u_i(\mu_i, \boldsymbol{\mu}_{-i}, \theta)}{\partial \mu_i}, \qquad i \in N \quad (13)$$

describes the long-run stochastic dynamics of the mean action of player i. Eq. 13 is a gradient dynamical system, where every player can be seen as maximizing his own utility function. In previous work, such gradient dynamics have been used as an *ad hoc* model of learning dynamics. It is interesting that we recover this result by explictly modeling the stochastic learning dynamics.

A counterintuitive feature of eq. 13, when we compare it to eq. 6, is that according to eq. 13, learning speed should increase with the standard deviation σ_i, while in the original stochastic updating rule (eq. 6), larger σ_i should lead to smaller updates to the mean action $\mu_{i,t+1}$.

It is noteworthy that at an equilibrium for the means $\hat{\boldsymbol{\mu}}$, action choice is still stochastic so that the equilibrium action of every player is $\hat{a}_i \sim \mathcal{N}(\hat{\mu}_i, \sigma_i)$.

5.3. **Simulations.**

TO DO: We will simulate the above game dynamics (eq. 6) in continuous-payoff games in order to compare it to our analysis based on stochastic approximation.

5.4. **Bounded action space.**

A difficulty that may arise from defining the chosen action as a normal random variable is that normal random variables are defined over the entire real line, $\mathbb{R}$. In certain cases, we may want to consider a bounded state space, of the form $(0, M)$. One way to achieve this is to make the transformation

$$a_i = \frac{M}{1 + \exp[-\tilde{a}_i]}, \tag{14}$$

where $\tilde{a}_i \in \mathbb{R}$ is the variable that is dynamically updated in eq. 6, and $a_i \in (0, M)$ is the action of individual i. For instance in a public-goods game, M may represent the endowment of each player, which is the maximum contribution he can make to the public good.

5.5. **Baseline simple individual decision problem.**

In this section we display simulations of the learning algorithm in the 1-player case under a deterministic environment. The utility function used here is given by

$$u_i(a_i) = -(a_i - \beta)^2, \tag{15}$$

which displays a single global optimum action achieved at $a_i = \beta$. Though there is nothing particularly exciting about this model, it is used only to show numerically that our stochastic approximation results presented above do hold and help predict the real dynamics of the players' actions.

TO DO: test in multi-player games with more complex payoff functions.

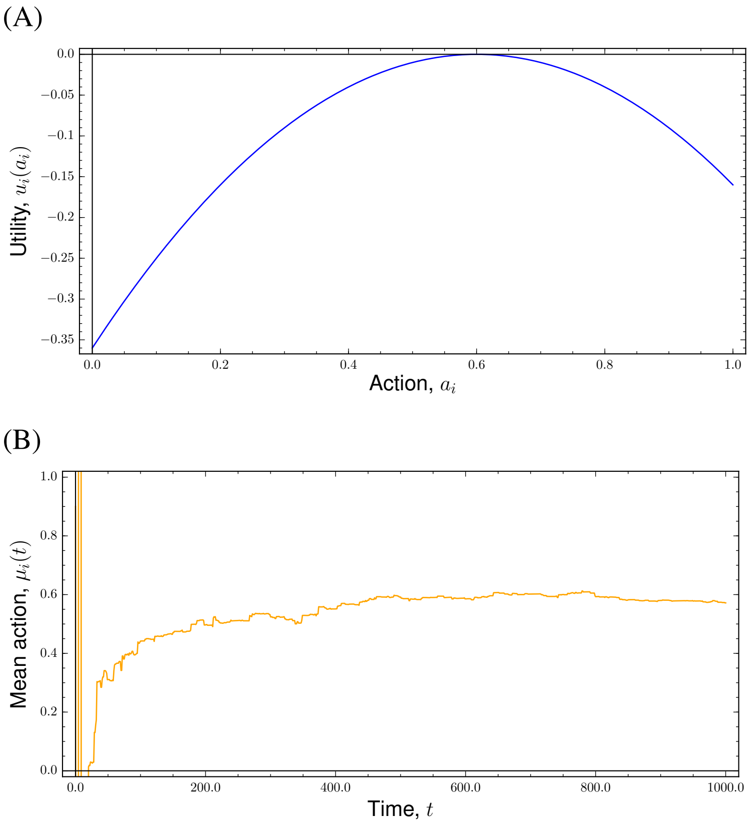

FIGURE 5. Continuous learning in a deterministic 1-player game. (A) The function $u_i(a_i)$ to be maximized by the player (eq. 15), with $\beta = 0.6$. (B) Typical dynamics of the mean action, $\mu_{i,t+1}$ when the utility function is that in (A).