Supplementary Information

Introspection dynamics in asymmetric multiplayers games

Marta Couto[1] and Saptarshi Pal[1]

[1]Max Planck Research Group Dynamics of Social Behavior, Max Planck Institute for Evolutionary Biology, 24306 Ploen, Germany

The general model for introspection dynamics

We consider a normal form game with $N(>2)$ players. Each player, i, has access to their action set, $\mathbf{A}^i$, in which there are m^i possible actions that they can play, $\mathbf{A}^i = \{a_1^i, a_2^i, ..., a_{m^i}^i\}$. The payoff for an individual i depends on what everyone plays and is represented by $\pi^i(\mathbf{a}^1, \mathbf{a}^2, ..., \mathbf{a}^N)$, where $\mathbf{a} := (\mathbf{a}^1, \mathbf{a}^2, ..., \mathbf{a}^N) \in \mathbf{A}^1 \times \mathbf{A}^2 \times ... \times \mathbf{A}^N =: \mathbf{A}$ represents a state of the game. In this model we only consider pure strategies for players. So, player i has only m^i possible strategies. Therefore, the total number of states possible for the game is the product of all m^i.

We represent a state of the game, $\mathbf{a}$, from the perspective of player j as $\mathbf{a} =: (\mathbf{a}^j, \mathbf{a}^{-j})$. In the state $(\mathbf{a}^j, \mathbf{a}^{-j})$, player j plays the action $\mathbf{a}^j \in \mathbf{A}^j$ and all the other co-players of j play $\mathbf{a}^{-j} \in \Pi_{k\neq j}\mathbf{A}^k$. The payoff of player j in this state is also represented as $\pi^j(\mathbf{a}^j, \mathbf{a}^{-j})$.

The players update their strategies over time using the introspection dynamics[1]. At every time step, one randomly chosen player can update their strategy. The randomly chosen player, say j, currently playing strategy a_k^j, compares their current payoff to the payoff that they would obtain if they played a randomly selected strategy, $a_l^j \neq a_k^j$, from their action set $\mathbf{A}^j$. This comparison is done while assuming that the co-players do not change from $\mathbf{a}^{-j}$. Then, they adopt this new strategy with the probability:

$$p^j_{a_k^j \to a_l^j}(\mathbf{a}^{-j}) = \frac{1}{1 + e^{-\beta(\pi^j(a_l^j, \mathbf{a}^{-j}) - \pi^j(a_k^j, \mathbf{a}^{-j}))}} \tag{1}$$

Where β is the selection parameter. The probability that no update is made by the randomly chosen player is given by the normalization condition:

$$p^j_{a_k^j \to a_k^j}(\mathbf{a}^{-j}) = 1 - \sum_{k\neq l} p^j_{a_k^j \to a_l^j}(\mathbf{a}^{-j}) \tag{2}$$

Now, with the help of Eq. 1 and 2, we can define the transition matrix $\mathbf{T}$ of the process. The transition

matrix is a square matrix of size $M = m^1 \times m^2 \times ... \times m^N$. The element $\mathbf{T}_{\mathbf{a}_p,\mathbf{a}_q}$ of the matrix represents the probability that the game will go to state $\mathbf{a}_q$ from state $\mathbf{a}_p$ in a time step. Before defining the transition matrix, we define the notion of neighbouring states.

Definition 1. ***Neighbouring states****:*
The state $\mathbf{a}_q \in Neb(\mathbf{a}_p)$*, the neighbourhood of state* $\mathbf{a}_p$*, if and only if* $\exists j$ *such that* $\mathbf{a}_p^{-j} = \mathbf{a}_q^{-j}$ *and* $\mathbf{a}_p^j \neq \mathbf{a}_q^j$.

Note that if $\mathbf{a}_q \in Neb(\mathbf{a}_p)$ then by definition $\mathbf{a}_p \in Neb(\mathbf{a}_q)$. We give an example to explain the notion of neighbourhood. Consider three players, player $1, 2$ and 3. Each of these players choose actions from an identical binary action set $\{0, 1\}$. In this example, the state $(0, 0, 0)$ is a neighbour of $(0, 0, 1)$ and vice versa. The index of difference between the two states is player 3. The states $(0, 0, 0)$ and $(0, 1, 1)$ are not neighbouring states because more than one player's action are changed.

Definition 2. ***Index of difference for neighbouring states****:*

If the states $\mathbf{a}_p$ *and* $\mathbf{a}_q$ *are neighbours then, the unique* j *for which* $\mathbf{a}_p^j \neq \mathbf{a}_q^j$*, is called the index of difference. Or* $I(\mathbf{a}_p, \mathbf{a}_q) := j$

For the earlier example, let $\mathbf{a}_p = (0, 0, 0)$ and $\mathbf{a}_q = (0, 0, 1)$. Then, $I(\mathbf{a}_p, \mathbf{a}_q) = 3$.

Using the above definitions of neighbourhood and index of difference, we can define the transition matrix $\mathbf{T}$ for the process as the follows:

$$\mathbf{T}_{\mathbf{a}_p,\mathbf{a}_q} = \begin{cases} \frac{1}{N(m_j-1)} \cdot p^j_{\mathbf{a}_p^j \to \mathbf{a}_q^j}(\mathbf{a}_p^{-j}) & \mathbf{a}_p \in Neb(\mathbf{a}_q) \quad \text{and,} \quad j = I(\mathbf{a}_p, \mathbf{a}_q) \\ 0 & \mathbf{a}_p \notin Neb(\mathbf{a}_q) \\ 1 - \sum_{\forall \mathbf{a}_k \neq \mathbf{a}_p} \mathbf{T}_{\mathbf{a}_p,\mathbf{a}_k} & \mathbf{a}_p = \mathbf{a}_q \end{cases} \tag{3}$$

The stochastic transition matrix $\mathbf{T}$ is row-stochastic. This implies that its stationary distribution is a left eigenvector of $\mathbf{T}$ corresponding to eigenvalue 1. We denote the stationary distribution of $\mathbf{T}$ as $\mathbf{u} := (\mathbf{u_a})_{\mathbf{a} \in \mathbf{A}}$. Furthermore, since no state of the process is isolated (i.e., every state has atleast one neighbour), a finite β assures that a state is reachable from any other state with positive probability in a finite number of steps. Therefore, $\mathbf{T}$ is also primitive. The stationary distribution $\mathbf{u}$ is unique and strictly positive (from the Perron-Frobenius Theorem). The following conditions are satisfied by the stationary distribution $\mathbf{u}$ when it is additionally also a probability distribution.

$$\mathbf{uT} = \mathbf{u} \tag{4}$$

$$\mathbf{u} \cdot \mathbf{1} = 1 \tag{5}$$

The marginal distribution over the strategies of a player j at the stationary distribution is given by $\xi^j := (\xi^j_{a^j_1}, \xi^j_{a^j_2}, ..., \xi^j_{a^j_{m_j}})$. This distribution indicates the probability with which player j plays strategies from their action set $\mathbf{A}^j$ at the stationary distribution. The relationship between the marginal distribution and the stationary distribution is given by:

$$\xi^j_{a^j_k} := \sum_{\mathbf{b} \in \mathbf{A}^{-j}} \mathbf{u}_{(a^j_k, \mathbf{b})} \tag{6}$$

Where, $\mathbf{A}^{-j} = \prod_{l \neq j} \mathbf{A}^l$. One can note that the marginal distribution is also a probability distribution. That is, for all j:

$$\sum_{a' \in \mathbf{A}^j} \xi^j_{a'} = 1 \tag{7}$$

Additive games and their properties under introspection dynamics

In this subsection we discuss a special class of general multiplayer games: additive games[2] and some of their properties under the introspection dynamics. Additive games are games where the payoff earned by a player j in the state $\mathbf{a}$ can be broken down into two-components - one which is only dependent on what player j played in the state: $\mathbf{a}^j$ and the other which only depends on what their co-players $-j$ played in the state: $\mathbf{a}^{-j}$. That is, for all players j,

$$\pi^j(\mathbf{a}) = \pi^j_{\mathbf{a}^j} + \pi^j_{\mathbf{a}^{-j}} \tag{8}$$

holds. In additive games, the payoff difference earned by a player when they switch between two strategies in their action set is independent of what other co-players are playing (given all other co-players do not change their strategies). That is, for all j and any pair of strategies $a^j_p, a^j_q \in \mathbf{A}^j$,

$$\pi^j(a^j_p, \mathbf{b}) - \pi^j(a^j_q, \mathbf{b}) = \pi^j_{a^j_p} - \pi^j_{a^j_q} \tag{9}$$

holds for any $\mathbf{b} \in \prod_{l \neq j} \mathbf{A}^l$.

We see an example of additive game in the next section. Under introspection dynamics, stationary distribution and marginal distributions from additive games show interesting properties. They are discussed in the theorems below:

Proposition 1. *When β is finite, the unique stationary distribution, $\mathbf{u} = (\mathbf{u_a})_{\mathbf{a}\in\mathbf{A}}$, of the introspection dynamics for the N-player additive game is given by:*

$$\mathbf{u}_a = \prod_{j=1}^{N} \frac{1}{\sum\limits_{a'\in\mathbf{A}^j} e^{\beta\left(\pi^j_{a'} - \pi^j_{\mathbf{a}^j}\right)}} \tag{10}$$

where, $\pi^j_{a'} - \pi^j_{\mathbf{a}^j}$ is the co-player independent payoff difference that j earns in an additive game when they unilaterally switch their play to a' from $\mathbf{a}^j$.

The above theorem gives a closed form expression of the stationary distribution of the introspection dynamics for additive multiplayer games. For any finite value of the selection parameter, β, the stationary distribution can be analytically calculated using the expression in Eq. (10). The expression is derived by assuming that the update process of the introspection dynamics is governed by Eq. (1) and (2). For proof, see Appendix .

Proposition 2. *Let $\mathbf{u} = (\mathbf{u_a})_{\mathbf{a}\in\mathbf{A}}$ be the unique stationary distribution of the introspection dynamics with finite β for the N-player additive game. Then, $\mathbf{u_a}$ is the product of the marginal probabilities that each player plays their respective actions in $\mathbf{a}$. That is,*

$$\mathbf{u_a} = \prod_{j=1}^{N} \xi^j_{\mathbf{a}^j} \tag{11}$$

where the marginal probability, $\xi^j_{\mathbf{a}^j}$, is the cumulative probability that player j plays $\mathbf{a}^j$ at the stationary distribution. The marginal probability is given by:

$$\xi^j_{\mathbf{a}_j} = \frac{1}{\sum\limits_{a'\in\mathbf{A}^j} e^{\beta\left(\pi^j_{a'} - \pi^j_{\mathbf{a}^j}\right)}} \tag{12}$$

where, $\pi^j_{a'} - \pi^j_{\mathbf{a}^j}$ is the co-player independent payoff difference that j earns in an additive game when they unilaterally switch their play to a' from $\mathbf{a}^j$ in an additive game.

In other words, the above theorem states for additive games, the probability that the game will be in state $\mathbf{a} = (\mathbf{a}^1, \mathbf{a}^2, ..., \mathbf{a}^N)$ in the long run (considering a finite selection strength introspection dynamics) is the product of the cumulative probabilities that each player j respectively adopts the action $\mathbf{a}^j$ in the stationary distribution.

To elucidate this result further, we use an example of a 2-player additive game where each player has three possible actions from the action set $\{0, 1, 2\}$. The result says that the expected frequency of the state $(0, 2)$, say, where player 1 and player 2 simultaneously play 0 and 2 respectively: $\mathbf{u}_{(0,2)}$, is the

product of the marginal probabilities that the first player plays 0 and the second player plays 2 in the game. That is,

$$\mathbf{u}_{(0,2)} = \xi_0^1 \xi_2^2 = (\mathbf{u}_{(0,0)} + \mathbf{u}_{(0,1)} + \mathbf{u}_{(0,2)})(\mathbf{u}_{(0,2)} + \mathbf{u}_{(1,2)} + \mathbf{u}_{(2,2)}) \tag{13}$$

As stated in the theorem above, this result holds for any additive game with arbitrary number of players and each player having arbitrary action sets.

Example of multiplayer additive game: linear public goods game

We consider the linear public goods game with N asymmetric players. Each player has two possible actions, to contribute (action, C), or to not contribute (action, D). The players differ in their cost of cooperation and their productivities. The cost of cooperation and the productivity of a player i are denoted by c^i and r^i. The payoff for player i when the state of the game is $\mathbf{a}$ is given by:

$$\pi^i(\mathbf{a}) = \sum_{j=1}^{N} \frac{r^j a^j c^j}{N} - a^i c^i \tag{14}$$

The difference in payoffs between playing the two actions in a linear public goods game is independent of what the other co-players play in the game. That is,

$$\pi^i(\mathbf{a}^i = \mathrm{D}, \mathbf{a}^{-i}) - \pi^i(\mathbf{a}^i = \mathrm{C}, \mathbf{a}^{-i}) = c^i \left(1 - \frac{r^i}{N}\right) =: f(c^i, r^i) \tag{15}$$

is independent of $\mathbf{a}^{-i}$. The linear public goods game is therefore an example of an additive game. This property of the game results in dominated strategies for player i depending on the relationship between their productivity r^i and number of players N. When $r^i < N$, the action of not contributing (D) dominates the action of contributing (C) and vice-versa.

Proposition 3. *When β is finite, the unique stationary distribution of the introspection dynamics for an asymmetric linear public goods game under introspection dynamics is given by:*

$$\mathbf{u}_{\mathbf{a}} = \prod_{j=1}^{N} \frac{1}{1 + e^{sign(\mathbf{a}^j)\beta f(c^j, r^j)}} \quad , \forall \mathbf{a} \in \{\mathrm{C}, \mathrm{D}\}^N \tag{16}$$

where,

$$sign(a) = \begin{cases} 1 & \text{if} \quad a = \mathrm{C} \\ -1 & \text{if} \quad a = \mathrm{D} \end{cases} \tag{17}$$

Symmetric multiplayer games with two strategies

In the previous section we looked at properties of the stationary distribution of the introspection dynamics for additive games. In this section we look at properties of the stationary distribution for any multiplayer game where players are symmetric and have two actions in their action set. A $N-$player symmetric normal form game satisfies the following properties:

1. $\mathbf{A}^1 = \mathbf{A}^2 = ... = \mathbf{A}^N = \mathcal{A}$, and
2. For any $i, j \in \{1, 2, ..., N\}, a \in \mathcal{A}$ and $\mathbf{b} \in \prod_{k=1}^{N-1} \mathcal{A}$:

$$\pi^i(a, \mathbf{b}) = \pi^j(a, \mathbf{b}) \tag{18}$$

In addition, when there are only two actions in the action set of players we define the action set in the same way as we did in the section of linear public goods game: $\mathcal{A} := \{\mathrm{C}, \mathrm{D}\}$.

Notations

We introduce some notations before presenting the results

1. We denote the payoffs of a C and D player when they have j co-players playing C, as $\pi^{\mathrm{C}}(j)$ and $\pi^{\mathrm{D}}(j)$ respectively.
2. We define a simple mapping function $\alpha(.)$ that maps the action C to 1 and the action D to 0. That is,

$$\alpha(a) := \begin{cases} 1 & \text{if} \quad a = \mathrm{C} \\ 0 & \text{if} \quad a = \mathrm{D} \end{cases} \tag{19}$$

3. For every state $\mathbf{a} \in \mathcal{A}^N$, we count the number of C players in the state and denote it with $\mathcal{C}(\mathbf{a})$. That is,

$$\mathcal{C}(\mathbf{a}) := \sum_{j=1}^{N} \alpha(\mathbf{a}^j) \tag{20}$$

Proposition 4. *When β is finite, the unique stationary distribution of the introspection dynamics for the $N-$player symmetric normal form game with two actions, $\mathcal{A} = \{\mathrm{C}, \mathrm{D}\}$, $(\mathbf{u_a})_{\mathbf{a}\in\mathcal{A}^N}$, is given by:*

$$\mathbf{u_a} = \frac{1}{\Gamma} \prod_{j=1}^{\mathcal{C}(\mathbf{a})} e^{-\beta f(j-1)} \tag{21}$$

where $f(j) := \pi^{\mathrm{D}}(j) - \pi^{\mathrm{C}}(j)$ is the difference in payoffs of playing D and C when there are j co-players playing C. The term Γ is the normalization factor given by:

$$\Gamma = \sum_{\forall \mathbf{a}' \in \mathcal{A}^N} \prod_{j=1}^{\mathcal{C}(\mathbf{a}')} e^{-\beta f(j-1)} \tag{22}$$

The above proposition gives an exact closed form expression for the steady state distribution of the introspection dynamics for symmetric games with two strategies. Since all players are symmetric in this two strategy game, the number of unique states of the game can be reduced to $N + 1$ from 2^N. Now, the state k corresponds to k players playing C and $N - k$ players playing D. Then, Proposition 4 can be simplified by accounting for states with number of C players as follows:

Corollary 1. *When β is finite, the unique stationary distribution, $(\mathbf{u}_k)_{k\in\{0,1,\ldots,N\}}$, of the introspection dynamics for the $N-$player symmetric normal form game with two actions, $\mathcal{A} = \{\mathrm{C}, \mathrm{D}\}$, is given by*

$$\mathbf{u}_k = \frac{1}{\Gamma} \cdot \binom{N}{k} \cdot \prod_{j=1}^{k} e^{-\beta f(j-1)} \tag{23}$$

where, k represents the number of C players in the state and $f(j) := \pi^{\mathrm{D}}(j) - \pi^{\mathrm{C}}(j)$ is the difference in payoffs of playing D and C when there are j co-players playing C. The term Γ is the normalization factor, given by,

$$\Gamma = \sum_{k=0}^{N} \binom{N}{k} \cdot \prod_{j=1}^{k+1} e^{-\beta f(j-1)} \tag{24}$$

The above lemma follows directly from Proposition 4. The key step is to count the number of states in the state space $\mathcal{A}^N$ that corresponds to exactly k, C players and $N - k$, D players. This count is simply the binomal coefficient $\binom{N}{k}$. For details see the proof in the Appendix.

Appendix: Proofs

Proof. **Proof of Proposition** 1

Since β is finite, the transition matrix of the process $\mathbf{T}$ given by Eq. 3 is primitive and therefore, the stationary distribution $\mathbf{u} = (\mathbf{u_a})_{\mathbf{a}\in\mathbf{A}}$ of the row-stochastic transition matrix is unique and satisfies the conditions laid out in Eq. 4 and 5. To continue for the rest of the proof, we introduce some short-cut notation that will be of use later in the proof:

$$I_q := I(\mathbf{a}_q, \mathbf{a}), \quad \textit{iff} \quad \mathbf{a}_q \in Neb(\mathbf{a}) \tag{25}$$

$$\tau^j_{\mathbf{a}^j} := \frac{1}{\sum\limits_{a'\in\mathbf{A}^j} e^{\beta\left(\pi^j_{a'} - \pi^j_{\mathbf{a}^j}\right)}} \tag{26}$$

In order to show that the candidate stationary distribution, as proposed in Eq. 10 is the stationary distribution of the process, we need to show that the following are true:

$$\mathbf{T}_{\mathbf{a},\mathbf{a}}\mathbf{u}_\mathbf{a} + \sum_{\mathbf{a}_q\neq\mathbf{a}} \mathbf{T}_{\mathbf{a}_q,\mathbf{a}}\mathbf{u}_{\mathbf{a}_q} = \mathbf{u}_a \quad \forall \mathbf{a} \in \mathbf{A} \tag{27}$$

$$\sum_{\mathbf{a}\in\mathbf{A}} \mathbf{u}_\mathbf{a} = 1 \tag{28}$$

The Eq 27 can be simplified further with the steps:

$$\mathbf{T}_{\mathbf{a},\mathbf{a}}\mathbf{u}_\mathbf{a} + \sum_{\mathbf{a}_q\neq\mathbf{a}} \mathbf{T}_{\mathbf{a}_q,\mathbf{a}}\mathbf{u}_{\mathbf{a}_q} = \tag{29}$$

$$\left(1 - \frac{1}{N}\sum_{\mathbf{a}_q\in Neb(\mathbf{a})} \frac{1}{m^{I_q}-1}\cdot p^{I_q}_{\mathbf{a}^{I_q}\to\mathbf{a}^{I_q}_q}\cdot\mathbf{u}_\mathbf{a}\right) + \sum_{\mathbf{a}_q\in Neb(\mathbf{a})} \frac{1}{m^{I_q}-1}\cdot p^{I_q}_{\mathbf{a}^{I_q}_q\to\mathbf{a}^{I_q}}\cdot\mathbf{u}_{\mathbf{a}_q} = \tag{30}$$

$$\mathbf{u}_\mathbf{a} + \frac{1}{N}\sum_{\mathbf{a}_q\in Neb(\mathbf{a})} \left(\prod_{k\neq I_q}\tau^k_{\mathbf{a}^k}\right)\left(p^{I_q}_{\mathbf{a}^{I_q}_q\to\mathbf{a}^{I_q}}\cdot\tau^{I_q}_{\mathbf{a}^{I_q}} - p^{I_q}_{\mathbf{a}^{I_q}\to\mathbf{a}^{I_q}_q}\cdot\tau^{I_q}_{\mathbf{a}^{I_q}_q}\right)\cdot\left(\frac{1}{m^{I_q}-1}\right) \tag{31}$$

Using the definition of probability of update of the introspection dynamics, as given by Eq. 1 and Eq. 26, it can be shown that:

$$p^{I_q}_{\mathbf{a}^{I_q}_q\to\mathbf{a}^{I_q}}\cdot\tau^{I_q}_{\mathbf{a}^{I_q}} - p^{I_q}_{\mathbf{a}^{I_q}\to\mathbf{a}^{I_q}_q}\cdot\tau^{I_q}_{\mathbf{a}^{I_q}_q} = 0 \tag{32}$$

Plugging the equality in Eq. 32 into Eq. 31, we can see that the left hand side of Eq. 27 indeed simplifies to $\mathbf{u_a}$. Now, to confirm that the candidate $\mathbf{u}$ is the unique stationary distribution we need to check if Eq. 28 holds. Simplifying the left hand side of this equation shows that:

$$\sum_{\mathbf{a}\in\mathbf{A}} \mathbf{u_a} = \sum_{\mathbf{a}\in\mathbf{A}} \prod_{k=1}^{N} \tau^k_{\mathbf{a}^k} \tag{33}$$

$$= \left(\prod_{k=1}^{N} \sum_{\mathbf{a'}\in\mathbf{A}} e^{\beta\pi^k_{\mathbf{a'}}}\right)^{-1} \cdot \left(\sum_{\mathbf{a}\in\mathbf{A}} \prod_{k=1}^{N} e^{\beta\pi^k_{\mathbf{a}^k}}\right) \tag{34}$$

$$= 1 \tag{35}$$

The step from Eq. 34 to Eq. 35 is possible because the sum and product in Eq. 34 are interchangeable for both the terms. Therefore, condition Eq. 28 is satisfied too. □

Proof. **Proof of Proposition** 2

If $\mathbf{u}$ is the unique stationary distribution of the $N-$player additive game with under the finite selection introspection dynamics, it is given by the expression in Eq. 10. We calculate the marginal distribution of any arbitrary state $\mathbf{a}$, $\xi_{\mathbf{a}} = (\xi^j_{\mathbf{a}^j})_{j=1,2,\ldots,N}$ by using the definition of marginal distribution in Eq. 6. It

follows that:

$$\xi^j_{\mathbf{a}^j} = \sum_{\mathbf{b}\in\mathbf{A}^{-j}} \mathbf{u}_{(\mathbf{a}^j,\mathbf{b})} \tag{36}$$

$$= \left(\prod_{k=1}^{N}\sum_{\mathbf{a}'\in\mathbf{A}^k} e^{\beta\pi^k_{\mathbf{a}'}}\right)^{-1} \cdot e^{\beta\pi^j_{\mathbf{a}^j}} \cdot \left(\sum_{\mathbf{b}\in\mathbf{A}^{-j}}\prod_{k\neq j} e^{\beta\pi^k_{\mathbf{b}^k}}\right) \tag{37}$$

$$= \left(\sum_{\mathbf{a}'\in\mathbf{A}^j} e^{\beta\pi^j_{\mathbf{a}'}}\right)^{-1} \cdot e^{\beta\pi^j_{\mathbf{a}^j}} \cdot \left(\prod_{k\neq j}\sum_{\mathbf{a}'\in\mathbf{A}^k} e^{\beta\pi^k_{\mathbf{a}'}}\right)^{-1} \cdot \left(\sum_{\mathbf{b}\in\mathbf{A}^{-j}}\prod_{k\neq j} e^{\beta\pi^k_{\mathbf{b}^k}}\right) \tag{38}$$

$$= \left(\sum_{\mathbf{a}'\in\mathbf{A}^j} e^{\beta\pi^j_{\mathbf{a}'}}\right)^{-1} \cdot e^{\beta\pi^j_{\mathbf{a}^j}} \cdot \left(\sum_{\mathbf{a}'\in\mathbf{A}^{-j}}\prod_{k\neq j} e^{\beta\pi^k_{\mathbf{a}'}}\right)^{-1} \cdot \left(\sum_{\mathbf{b}\in\mathbf{A}^{-j}}\prod_{k\neq j} e^{\beta\pi^k_{\mathbf{b}^k}}\right) \tag{39}$$

$$= \left(\sum_{\mathbf{a}'\in\mathbf{A}^j} e^{\beta\left(\pi^j_{\mathbf{a}'}-\pi^j_{\mathbf{a}^j}\right)}\right)^{-1} \tag{40}$$

Therefore, the marginal distribution follows Eq. 12. Now since we also additionally know that the stationary distribution follows the form Eq. 10, we can conclude that for additive games, under introspection dynamics with finite selection, Eq. 11 holds. □

Proof. **Proof of Proposition** 3

Since we have demonstrated that the linear public goods game is an additive game, the proof of this theorem can be performed by directly using Theorem 1. Here, we provide an independent proof. The idea behind this proof is identical to the proof of Theorem 1.

The stationary transition matrix $\mathbf{T}$ for the linear public goods game is primitive when β is finite (i.e., there is a positive power k such that $\mathbf{T}^k$ is a strictly positive matrix). Therefore, the stationary distribution of $\mathbf{T}$ will always be unique. We define the following short cut notations for the ease of the proof:

$$\bar{\mathbf{a}}^j := \{\mathrm{D},\mathrm{C}\} \setminus \mathbf{a}^j \tag{41}$$

$$p^j := \frac{1}{1+\, e^{\beta f(c^j,r^j)}} \tag{42}$$

In addition we introduce a mapping function $\alpha(.)$ which maps the action C to 1 and the action D to 0.

That is $\alpha(\mathrm{C}) := 1$ and $\alpha(\mathrm{D}) := 0$. Using these notations and Eq. 1 and 15 we can write the probability that a player j updates from $\mathbf{a}^j$ to $\bar{\mathbf{a}}^j$ while their co-players play $\mathbf{a}^{-j}$ as:

$$p^j_{\bar{\mathbf{a}}^j \rightarrow \mathbf{a}^j}(\mathbf{a}^{-j}) = p^j sign(\mathbf{a}^j) + \alpha(\bar{\mathbf{a}}^j) \tag{43}$$

The candidate stationary distribution $\mathbf{u}$ given in Eq 16 can be written down using our shortcut notation as:

$$\mathbf{u}_{\mathbf{a}} = \prod_{k=1}^{N} p^k sign(\mathbf{a}^k) + \alpha(\bar{\mathbf{a}}^k) \quad , \forall \mathbf{a} \in \{0,1\}^N \tag{44}$$

This stationary distribution must satisfy the following properties, which are also given in Eq 4 and 5:

$$\mathbf{u}_{\mathbf{a}} = \mathbf{T}_{\mathbf{a},\mathbf{a}}\mathbf{u}_{\mathbf{a}} + \sum_{\mathbf{a}_q \neq \mathbf{a}} \mathbf{T}_{\mathbf{a}_q,\mathbf{a}}\mathbf{u}_{\mathbf{a}_q} \tag{45}$$

$$\sum_{\forall \mathbf{a}_q} \mathbf{u}_{\mathbf{a}_q} = 1 \tag{46}$$

Where, the terms in the right hand side of Eq. 45 can be simplified using Eq 1 and 3 as follows:

$$\mathbf{T}_{\mathbf{a},\mathbf{a}} = 1 - \sum_{k=1}^{N} \mathbf{T}_{(\mathbf{a}^k,\mathbf{a}^{-k}),(\bar{\mathbf{a}}^k,\mathbf{a}^{-k})} = 1 - \frac{1}{N}\sum_{k=1}^{N} p^k sign(\bar{\mathbf{a}}^k) + \alpha(\mathbf{a}^k) \tag{47}$$

and additionally, also using Eq. 44 the second term can be simplified too:

$$\sum_{\mathbf{a}_q \neq \mathbf{a}} \mathbf{T}_{\mathbf{a}_q,\mathbf{a}}\mathbf{u}_{\mathbf{a}_q} = \sum_{k=1}^{N} \mathbf{T}_{(\bar{\mathbf{a}}^k,\mathbf{a}^{-k}),(\mathbf{a}^k,\mathbf{a}^{-k})}\mathbf{u}_{(\bar{\mathbf{a}}^k,\mathbf{a}^{-k})} \tag{48}$$

$$= \frac{1}{N}\sum_{k=1}^{N} \left(p^k sign(\mathbf{a}^k) + \alpha(\bar{\mathbf{a}}^k)\right) \mathbf{u}_{(\bar{\mathbf{a}}^k,\mathbf{a}^{-k})} \tag{49}$$

$$= \frac{\mathbf{u}_{\mathbf{a}}}{N}\sum_{k=1}^{N} p^k sign(\bar{\mathbf{a}}^k) + \alpha(\mathbf{a}^k) \tag{50}$$

Now, using Eq. 47, 50 one can show that the right hand side of Eq. 45 is the element of the stationary distribution, corresponding to the state $\mathbf{a}$, $\mathbf{u}_a$. Now, to complete the proof, we must show that Eq. 46 is also true for our candidate stationary distribution. This can be done by decomposing the sum of the

elements of the stationary distribution as follows:

$$\sum_{\forall \mathbf{a}_q} u_{\mathbf{a}_q} = \sum_{\forall \mathbf{a}_q} \prod_{k=1}^{N} p^k sign(\mathbf{a}_q^k) + \alpha(\bar{\mathbf{a}}_q^k) \tag{51}$$

$$= \sum_{\forall \mathbf{a}_q} (1 - p^N) \prod_{k=1}^{N-1} p^k sign(\mathbf{a}_q^k) + \alpha(\bar{\mathbf{a}}_q^k) + p^N \prod_{k=1}^{N-1} p^k sign(\mathbf{a}_q^k) + \alpha(\bar{\mathbf{a}}_q^k) \tag{52}$$

$$= \sum_{\forall \mathbf{a}_q} \prod_{k=1}^{N-1} p^k sign(\mathbf{a}_q^k) + \alpha(\bar{\mathbf{a}}_q^k) \tag{53}$$

When the above decomposition is perfomed $N-1$ more times, the sum of the right hand side becomes 1. This prooves that the candidate stationary distribution is also a probability distribution. □

Proof. **Proof of Proposition** 4

By construction, the candidate stationary distribution given by Eq. 21 and Eq. 22 is a probability distribution since it satisfies the condition in Eq. 5 and for any state $\mathbf{a}'$, $\mathbf{u}_{\mathbf{a}'}$ is between 0 and 1. Moreover, since β is finite, the transition matrix of the process $\mathbf{T}$ is primitive and therefore, it will have a unique stationary distribution. To show that the candidate stationary distribution is the unique stationary distribution, we need to check if for this process, $\mathbf{uT} = \mathbf{u}$. That is, the Eq. 45 must hold. We re-introduce some notations that we will use in this proof:

$$\bar{\mathbf{a}}^j := \{\mathrm{D}, \mathrm{C}\} \setminus \mathbf{a}^j \tag{54}$$

$$\alpha(a) := \begin{cases} 1 & \text{if} \quad a = \mathrm{C} \\ \\ 0 & \text{if} \quad a = \mathrm{D} \end{cases} \tag{55}$$

$$\mathcal{C}(\mathbf{a}) = \sum_{j=1}^{N} \alpha(\mathbf{a}^j) \tag{56}$$

For this process, the first term in the right hand side of Eq. 45 can be simplified as:

$$\mathbf{u_a}\mathbf{T}_{\mathbf{a},\mathbf{a}} = \mathbf{u_a} - \mathbf{u_a} \sum_{k=1}^{N} \mathbf{T}_{(\mathbf{a}^k,\mathbf{a}^{-k}),(\bar{\mathbf{a}}^k,\mathbf{a}^{-k})} \tag{57}$$

$$= \mathbf{u_a} - \frac{\mathbf{u_a}}{N} \sum_{k=1}^{N} \frac{1}{1 + e^{sign(\bar{\mathbf{a}}^k)\beta f(N_k)}} \tag{58}$$

Where, the function $sign(.)$ is defined in Eq. 17 and $f(j)$ is the difference in payoffs between playing D and C when there are j co-players playing C. The term N_k is the number of co-players of k that play C in state $\mathbf{a}_{\mathrm{C}}$. That is,

$$N_k := \sum_{j \neq k} \alpha(\mathbf{a}^l) \tag{59}$$

The second term in the right hand side of Eq. 45 can be simplified as,

$$\sum_{\mathbf{a}_q \neq \mathbf{a}} \mathbf{T}_{\mathbf{a}_q,\mathbf{a}} \mathbf{u}_{\mathbf{a}_q} = \sum_{k=1}^{N} \mathbf{T}_{(\bar{\mathbf{a}}^k,\mathbf{a}^{-k}),(\mathbf{a}^k,\mathbf{a}^{-k})} \mathbf{u}_{(\bar{\mathbf{a}}^k,\mathbf{a}^{-k})} \tag{60}$$

$$= \frac{1}{N\Gamma} \sum_{k=1}^{N} \mathbf{T}_{(\bar{\mathbf{a}}^k,\mathbf{a}^{-k}),(\mathbf{a}^k,\mathbf{a}^{-k})} \prod_{j=1}^{\mathcal{C}((\bar{\mathbf{a}}^k,\mathbf{a}^{-k}))} e^{-\beta f(j-1)} \tag{61}$$

$$= \frac{1}{N\Gamma} \sum_{k=1}^{N} \mathbf{T}_{(\bar{\mathbf{a}}^k,\mathbf{a}^{-k}),(\mathbf{a}^k,\mathbf{a}^{-k})} \left(\prod_{j=1}^{\sum_{j \neq k} \alpha(\mathbf{a}^l)} e^{-\beta f(j-1)} \right) \cdot e^{-\beta \alpha(\bar{\mathbf{a}}^k) f(-\alpha(\mathbf{a}^k) + \sum_{j \neq k} \mathbf{a}^l)} \tag{62}$$

Between the steps, Eq. 61 and 62, we took out one term from the product that is present in our candidate distribution. This term accounts for the k^{th} players action in the state $(\bar{\mathbf{a}}^k, \mathbf{a}^{-k})$. For simplicity, we replace $\mathbf{T}_{(\bar{\mathbf{a}}^k,\mathbf{a}^{-k}),(\mathbf{a}^k,\mathbf{a}^{-k})}$ with just $\mathbf{T}$ in the next steps and also denote the number of co-players of k that play C in the state $\mathbf{a}$ as N_k. We continue the simplification of Eq. 62 in the later steps by introducing terms that cancel each other.

$$\sum_{\mathbf{a}_q \neq \mathbf{a}} \mathbf{T}_{\mathbf{a}_q,\mathbf{a}} \mathbf{u}_{\mathbf{a}_q} = \frac{1}{N\Gamma} \sum_{k=1}^{N} \mathbf{T} \cdot \left(\prod_{j=1}^{N_k} e^{-\beta f(j-1)} \right) \cdot \frac{e^{-\beta \alpha(\bar{\mathbf{a}}^k) f(-\alpha(\mathbf{a}^k) + N_k)}}{e^{-\beta \alpha(\mathbf{a}^k) f(-\alpha(\bar{\mathbf{a}}^k) + N_k)}} \cdot e^{-\beta \alpha(\mathbf{a}^k) f(-\alpha(\bar{\mathbf{a}}^k) + N_k)} \tag{63}$$

The newly introduced term in Eq. 63 can be taken inside the product. Note that this term is 0 if the k^{th} player plays D in the state $\mathbf{a}$. When this term is taken inside the product bracket, products of exponent $e^{-\beta f(j-1)}$ can be performed where j ranges from 1 to the number of cooperators in state $\mathbf{a}$, $\mathcal{C}(\mathbf{a})$. This

product is then the stationary distribution probability $\mathbf{u_a}$. That is,

$$\sum_{\mathbf{a}_q \neq \mathbf{a}} \mathbf{T}_{\mathbf{a}_q,\mathbf{a}} \mathbf{u}_{\mathbf{a}_q} = \frac{1}{N\Gamma} \sum_{k=1}^{N} \mathbf{T} \cdot \left(\prod_{j=1}^{N_k} e^{-\beta f(j-1)} \cdot e^{-\beta\alpha(\mathbf{a}^k) f(-\alpha(\bar{\mathbf{a}}^k)+N_k)} \right) \cdot \frac{e^{-\beta\alpha(\bar{\mathbf{a}}^k) f(-\alpha(\mathbf{a}^k)+N_k)}}{e^{-\beta\alpha(\mathbf{a}^k) f(-\alpha(\bar{\mathbf{a}}^k)+N_k)}} \tag{64}$$

$$= \frac{1}{N} \sum_{k=1}^{N} \mathbf{T} \cdot \left(\frac{1}{\Gamma} \prod_{j=1}^{\mathcal{C}(\mathbf{a})} e^{-\beta f(j-1)} \right) \cdot \frac{e^{-\beta\alpha(\bar{\mathbf{a}}^k) f(-\alpha(\mathbf{a}^k)+N_k)}}{e^{-\beta\alpha(\mathbf{a}^k) f(-\alpha(\bar{\mathbf{a}}^k)+N_k)}} \tag{65}$$

$$= \frac{1}{N} \sum_{k=1}^{N} \mathbf{T}_{(\bar{\mathbf{a}}^k,\mathbf{a}^{-k}),(\mathbf{a}^k,\mathbf{a}^{-k})} \cdot \mathbf{u_a} \cdot \frac{e^{-\beta\alpha(\bar{\mathbf{a}}^k) f(-\alpha(\mathbf{a}^k)+N_k)}}{e^{-\beta\alpha(\mathbf{a}^k) f(-\alpha(\bar{\mathbf{a}}^k)+N_k)}} \tag{66}$$

The fraction inside the sum in Eq. 66 can be simplified as follows leading to further simplification of Eq. 66:

$$\sum_{\mathbf{a}_q \neq \mathbf{a}} \mathbf{T}_{\mathbf{a}_q,\mathbf{a}} \mathbf{u}_{\mathbf{a}_q} = \frac{1}{N} \sum_{k=1}^{N} \mathbf{T}_{(\bar{\mathbf{a}}^k,\mathbf{a}^{-k}),(\mathbf{a}^k,\mathbf{a}^{-k})} \cdot \mathbf{u_a} \cdot e^{sign(\mathbf{a}^k)\beta f(N_k)} \tag{67}$$

In Eq. 67 we can replace the element of the transition matrix $\mathbf{T}_{(\bar{\mathbf{a}}^k,\mathbf{a}^{-k}),(\mathbf{a}^k,\mathbf{a}^{-k})}$ by using the following:

$$\mathbf{T}_{(\bar{\mathbf{a}}^k,\mathbf{a}^{-k}),(\mathbf{a}^k,\mathbf{a}^{-k})} = \frac{1}{1 + e^{sign(\mathbf{a}^k)\beta f(N_k)}} \tag{68}$$

Using the expression for the transition matrix element from Eq. 68 into Eq. 67 and by using Eq. 58, we can simplify further:

$$\sum_{\mathbf{a}_q \neq \mathbf{a}} \mathbf{T}_{\mathbf{a}_q,\mathbf{a}} \mathbf{u}_{\mathbf{a}_q} = \frac{\mathbf{u_a}}{N} \sum_{k=1}^{N} \frac{1}{1 + e^{sign(\mathbf{a}^k)\beta f(N_k)}} \cdot e^{sign(\mathbf{a}^k)\beta f(N_k)} \tag{69}$$

$$= \frac{\mathbf{u_a}}{N} \sum_{k=1}^{N} \frac{1}{1 + e^{sign(\bar{\mathbf{a}}^k)\beta f(N_k)}} \tag{70}$$

$$= \mathbf{u_a} - \mathbf{u_a}\mathbf{T}_{\mathbf{a},\mathbf{a}} \tag{71}$$

The final step in the previous simplification shows that Eq. 45 holds for any $\mathbf{a} \in \{\mathrm{C},\mathrm{D}\}^N$. Therefore, the candidate distribution we propose in Eq. 21 is the unique stationary distribution of the symmetric N-player game with two strategies. □

Proof. **Proof of Corollary** 1

□

Supplementary References

References

[1] Marta Couto, Stefano Giaimo, and Christian Hilbe. Introspection dynamics: A simple model of counterfactual learning in asymmetric games. *New Journal of Physics*, 2022.

[2] Alex McAvoy and Christoph Hauert. Asymmetric evolutionary games. *PLoS computational biology*, 11(8):e1004349, 2015.