Supplementary Information for "Super-additive cooperation"

Charles Efferson[1,*], Helen Bernhard[2], Urs Fischbacher[3,4], and Ernst Fehr[2]

[1]*Faculty of Business and Economics, University of Lausanne, Switzerland*

[2]*Department of Economics, University of Zurich, Switzerland*

[3]*Department of Economics, University of Konstanz, Germany*

[4]*Thurgau Institute of Economics, Switzerland*

[*]*Address correspondence to CE (charles.efferson@unil.ch).*

Contents

1 A sequential two-person social dilemma with a continuous action space

When two individuals are paired, they play a sequential social dilemma some number of times. In particular, each round of play is a single interaction, and interactions are indexed by $n \in \{1, 2, \ldots\}$. For any given interaction n, the interaction $n + 1$ occurs with probability $\delta \in [0, 1)$. As explained below, we switch to a fixed number of interactions per pair for our agent-based simulation models (§ 2). For now, however, we use a constant continuation probability for analytical results.

Each individual in a pair plays the role of first mover with probability 0.5. Once assigned, roles are fixed. For each interaction, each player has an endowment, which we normalise to 1. The first mover transfers $x_n \in [0, 1]$, and after learning the first mover's choice the second mover back transfers $y_n \in [0, 1]$. Transfers and back transfers yield efficiency gains. Specifically, for some $b > 1$, endogenous payoffs from interaction n are, for the first and second movers respectively,

$$\begin{aligned} \pi_n^1 &= 1 - x_n + by_n \\ \pi_n^2 &= 1 - y_n + bx_n. \end{aligned} \tag{1}$$

1.1 Two-dimensional strategies

A player's strategy involves two heritable quantities, namely the initial transfer if first mover and a second quantity that controls the player's response function. When appropriate, we will generally use i to index individuals and j to index groups, although at various times we will omit one or both indices. We will always clarify notation when we do so.

For the present analysis we omit j and arbitrarily call the first mover i and the second mover i'. Let $\tilde{x}_i \in [0, 1]$ be the initial transfer for i, and let $q_i \in [-1, 1]$ control i's response function. If $q_i > 0$, agent i is an escalating reciprocator (Fig. 1a), with $q_i = 1$ the extreme case in which i is unconditionally cooperative. If $q_i < 0$, agent i is a de-escalating reciprocator (Fig. 1b), with

$q_i = -1$ the extreme case in which i is unconditionally selfish. If $q_i = 0$, agent i is a perfect reciprocator, which is of course a degenerate case of both escalating and de-escalating reciprocity. Agent i' also has a heritable initial transfer, $\tilde{x}_{i'} \in [0, 1]$. As second mover in a pair, however, this quantity does not come into play. Instead, $q_{i'} \in [-1, 1]$ is the only component of the strategy of i' that affects play in the pair. To simplify the analyses below, we define the following. If $q_i \in [0, 1]$, $\alpha_i \coloneqq q_i$. If $q_i \in [-1, 0]$, $\psi_i \coloneqq 1 + q_i$. If $q_{i'} \in [0, 1]$, $\alpha_{i'} \coloneqq q_{i'}$. If $q_{i'} \in [-1, 0]$, $\psi_{i'} \coloneqq 1 + q_{i'}$.

1.1.1 Both players escalate

We begin by assuming both players in the pair escalate, perhaps weakly, their transfers. This means $q_i \in [0, 1]$, $\alpha_i = q_i$, $q_{i'} \in [0, 1]$, and $\alpha_{i'} = q_{i'}$. Altogether, strategies take the form,

$$\begin{aligned} n = 1, \quad & x_n = \tilde{x}_i \\ & y_n = (1 - \alpha_{i'})\tilde{x}_i + \alpha_{i'} \\ n \geq 2, \quad & x_n = (1 - \alpha_i)y_{n-1} + \alpha_i \\ & y_n = (1 - \alpha_{i'})x_n + \alpha_{i'}. \end{aligned} \tag{2}$$

The general solution for player choices is

$$\begin{aligned} n = 1, \quad & x_n = \tilde{x}_i \\ & y_n = (1 - \alpha_{i'})\tilde{x}_i + \alpha_{i'} \\ n \geq 2, \quad & x_n = (1 - \alpha_i)^{n-1}(1 - \alpha_{i'})^{n-1}\tilde{x}_i + \{\alpha_i + (1 - \alpha_i)\alpha_{i'}\} \sum_{k=0}^{n-2}(1 - \alpha_i)^k(1 - \alpha_{i'})^k \\ & y_n = (1 - \alpha_i)^{n-1}(1 - \alpha_{i'})^n\tilde{x}_i + \alpha_i(1 - \alpha_{i'}) \sum_{k=0}^{n-2}(1 - \alpha_i)^k(1 - \alpha_{i'})^k \\ & \qquad + \alpha_{i'} \sum_{k=0}^{n-1}(1 - \alpha_i)^k(1 - \alpha_{i'})^k. \end{aligned} \tag{3}$$

If $\alpha_i = 0$ and $\alpha_{i'} = 0$, both players are perfect reciprocators in the sense that, $\forall n \geq 1$, $x_n = y_n = \tilde{x}_i$. The situation is different, however, if at least one player is not a perfect reciprocator. Namely,

if $\alpha_i > 0$ or $\alpha_{i'} > 0$, for $l \in \{1, 2\}$,

$$\sum_{k=0}^{n-l}(1-\alpha_i)^k(1-\alpha_{i'})^k = \left(1-(1-\alpha_i)^{n-l+1}(1-\alpha_{i'})^{n-l+1}\right) / \left(1-(1-\alpha_i)(1-\alpha_{i'})\right). \quad (4)$$

As a result,

$$\begin{aligned} \lim_{n\to\infty} x_n &= \frac{\alpha_i + (1-\alpha_i)\alpha_{i'}}{1-(1-\alpha_i)(1-\alpha_{i'})} = 1 \\ \lim_{n\to\infty} y_n &= \frac{\alpha_i(1-\alpha_{i'}) + \alpha_{i'}}{1-(1-\alpha_i)(1-\alpha_{i'})} = 1. \end{aligned} \quad (5)$$

This tells us that, if at least one of the two players strictly escalates, they will jointly converge on full cooperation as their relationship continues.

1.1.2 Both players de-escalate

Now we assume both players in the pair de-escalate, perhaps weakly, their transfers. For the first mover, $\tilde{x}_i \in [0, 1]$ is the initial transfer for $n = 1$, $q_i \in [-1, 0]$ controls the response function, and $\psi_i = 1 + q_i$. The second mover's response function depends on $q_{i'} \in [-1, 0]$, which in turn means that $\psi_{i'} = 1 + q_{i'}$. Altogether, strategies take the form,

$$\begin{aligned} n = 1, \quad x_n &= \tilde{x}_i \\ y_n &= \psi_{i'}\tilde{x}_i \\ n \geq 2, \quad x_n &= \psi_i y_{n-1} \\ y_n &= \psi_{i'} x_n. \end{aligned} \quad (6)$$

For a pair of this sort, the general solution for their choices is

$$\begin{aligned} n \geq 1, \quad x_n &= \psi_i^{n-1}\psi_{i'}^{n-1}\tilde{x}_i \\ y_n &= \psi_i^{n-1}\psi_{i'}^{n}\tilde{x}_i. \end{aligned} \quad (7)$$

If $\psi_i = 1$ and $\psi_{i'} = 1$, both players are perfect reciprocators in the sense that, $\forall n \geq 1$, $x_n = y_n = \tilde{x}_i$. However, if $\psi_i < 1$ or $\psi_{i'} < 1$,

$$\begin{aligned} \lim_{n\to\infty} x_n &= 0 \\ \lim_{n\to\infty} y_n &= 0. \end{aligned} \tag{8}$$

This result tells us that, if at least one of the two players strictly de-escalates, they will jointly converge on full defection as their relationship continues.

1.1.3 First mover escalates, second mover de-escalates

For the first mover, $\tilde{x}_i \in [0, 1]$ is the initial transfer for $n = 1$, and $\alpha_i = q_i \in [0, 1]$ controls the response function. The second mover's response function depends on $\psi_{i'} = 1 + q_{i'}$, where $q_{i'} \in [-1, 0]$. Altogether, strategies take the form,

$$\begin{aligned} n = 1, \quad & x_n = \tilde{x}_i \\ & y_n = \psi_{i'}\tilde{x}_i \\ n \geq 2, \quad & x_n = (1 - \alpha_i)y_{n-1} + \alpha_i \\ & y_n = \psi_{i'}x_n. \end{aligned} \tag{9}$$

For a pair of this sort, the general solution for their choices is

$$\begin{aligned} n = 1, \quad & x_n = \tilde{x}_i \\ & y_n = \psi_{i'}\tilde{x}_i \\ n \geq 2, \quad & x_n = (1 - \alpha_i)^{n-1}\psi_{i'}^{n-1}\tilde{x}_i + \alpha_i \sum_{k=0}^{n-2}(1 - \alpha_i)^k\psi_{i'}^k \\ & y_n = (1 - \alpha_i)^{n-1}\psi_{i'}^n\tilde{x}_i + \alpha_i\psi_{i'} \sum_{k=0}^{n-2}(1 - \alpha_i)^k\psi_{i'}^k. \end{aligned} \tag{10}$$

If $\alpha_i = 0$ and $\psi_{i'} = 1$, both players are perfect reciprocators in the sense that, $\forall n \geq 1$, $x_n = y_n = \tilde{x}_i$. If $\alpha_i > 0$ or $\psi_{i'} < 1$,

$$\sum_{k=0}^{n-2}(1 - \alpha_i)^k\psi_{i'}^k = \left(1 - (1 - \alpha_i)^{n-1}\psi_{i'}^{n-1}\right) / \left(1 - (1 - \alpha_i)\psi_{i'}\right). \tag{11}$$

In this case,

$$\begin{aligned}\lim_{n\to\infty} x_n &= \frac{\alpha_i}{1-(1-\alpha_i)\psi_{i'}} \\ \lim_{n\to\infty} y_n &= \frac{\alpha_i\psi_{i'}}{1-(1-\alpha_i)\psi_{i'}}.\end{aligned} \tag{12}$$

This result tells us the following. If the first mover strictly escalates ($\alpha_i > 0$) and the second mover perfectly reciprocates ($\psi_{i'} = 1$), the players converge on full cooperation as their relationship continues. If the first mover perfectly reciprocates ($\alpha_i = 0$) and the second mover strictly de-escalates ($\psi_{i'} < 1$), the players converge on full defection as their relationship continues. If the first mover strictly escalates ($\alpha_i > 0$) and the second mover strictly de-escalates ($\psi_{i'} < 1$), the first mover converges to a contribution of $\alpha_i / (1-(1-\alpha_i)\psi_{i'})$, where $0 < \alpha_i / (1-(1-\alpha_i)\psi_{i'}) \leq 1$. The second mover converges to a contribution of $\alpha_i\psi_{i'} / (1-(1-\alpha_i)\psi_{i'})$. Because $\psi_{i'} < 1$, the second mover converges to a contribution that is strictly less than the contribution to which the first mover converges. In this sense, the de-escalating second mover will consistently exploit the escalating first mover under a long-term relationship.

1.1.4 First mover de-escalates, second mover escalates

For the first mover, $\tilde{x}_i \in [0,1]$ is the initial transfer for $n = 1$, and $\psi_i = 1 + q_i$ controls the response function, where $q_i \in [-1, 0]$. The second mover's response function depends on $\alpha_{i'} = q_{i'} \in [0,1]$. Altogether, strategies take the form,

$$\begin{aligned} n = 1, \quad & x_n = \tilde{x}_i \\ & y_n = (1-\alpha_{i'})\tilde{x}_i + \alpha_{i'} \\ n \geq 2, \quad & x_n = \psi_i y_{n-1} \\ & y_n = (1-\alpha_{i'})x_n + \alpha_{i'}. \end{aligned} \tag{13}$$

For a pair of this sort, the general solution for their choices is

$$\begin{aligned} n = 1, \quad & x_n = \tilde{x}_i \\ & y_n = (1-\alpha_{i'})\tilde{x}_i + \alpha_{i'} \\ n \geq 2, \quad & x_n = \psi_i^{n-1}(1-\alpha_{i'})^{n-1}\tilde{x}_i + \psi_i \alpha_{i'} \sum_{k=0}^{n-2} \psi_i^k (1-\alpha_{i'})^k \\ & y_n = \psi_i^{n-1}(1-\alpha_{i'})^{n}\tilde{x}_i + \alpha_{i'} \sum_{k=0}^{n-1} \psi_i^k (1-\alpha_{i'})^k. \end{aligned} \tag{14}$$

If $\psi_i = 1$ and $\alpha_{i'} = 0$, both players are perfect reciprocators in the sense that, $\forall n \geq 1$, $x_n = y_n = \tilde{x}_i$. If $\psi_i < 1$ or $\alpha_{i'} > 0$, at least one player is not a perfect reciprocator. In this case one can show that, for $l \in \{1, 2\}$, $\sum_{k=0}^{n-l} \psi_i^k (1-\alpha_{i'})^k = \left(1 - \psi_i^{n-l+1}(1-\alpha_{i'})^{n-l+1}\right) / \left(1 - \psi_i(1-\alpha_{i'})\right)$. Choices converge as,

$$\begin{aligned} \lim_{n\to\infty} x_n &= \frac{\psi_i \alpha_{i'}}{1 - \psi_i(1-\alpha_{i'})} \\ \lim_{n\to\infty} y_n &= \frac{\alpha_{i'}}{1 - \psi_i(1-\alpha_{i'})}. \end{aligned} \tag{15}$$

This result tells us the following. If the first mover perfectly reciprocates ($\psi_i = 1$) and the second mover strictly escalates ($\alpha_{i'} > 0$), the players converge on full cooperation as their relationship continues. If the first mover strictly de-escalates ($\psi_i < 1$) and the second mover perfectly reciprocates ($\alpha_{i'} = 0$), the players converge on full defection. If the first mover strictly de-escalates ($\psi_i < 1$) and the second mover strictly escalates ($\alpha_{i'} > 0$), the first mover converges to a contribution $\psi_i \alpha_{i'} / \left(1 - \psi_i(1-\alpha_{i'})\right)$, where $0 \leq \psi_i \alpha_{i'} / \left(1 - \psi_i(1-\alpha_{i'})\right) < 1$. The second mover converges to a contribution of $\alpha_{i'} / \left(1 - \psi_i(1-\alpha_{i'})\right)$, where $0 \leq \psi_i \alpha_{i'} / \left(1 - \psi_i(1-\alpha_{i'})\right) < \alpha_{i'} / \left(1 - \psi_i(1-\alpha_{i'})\right) \leq 1$. In this sense the de-escalating first mover will consistently exploit the escalating second mover under a long-term relationship.

1.1.5 An analysis of de-escalating monomorphisms

An analysis of de-escalating monomorphisms requires us to change our perspective. Instead of considering a pair that has already formed, with i the first mover and i' the second mover, we now have to consider the fitness values of a resident strategy, D, and a rare mutant, D', that appears at some point in time. Let $\Pi_1(D;D)$ be a random variable with support $\mathbb{R}_+$ that denotes the fitness of a de-escalating first mover of the resident type (i.e. the argument D) with a de-escalating partner having the same strategy (i.e. the exogenous D). Similarly, $\Pi_2(D;D)$ is the equivalent random variable for the fitness of a de-escalating second mover of the resident type with a partner having the same strategy. Because we are examining rare mutants in an otherwise monomorphic population, we drop the i subscripts.

Individuals of the resident type are characterised by initial transfers, $\tilde{x}$, and a degree of de-escalation $\psi \in [0,1]$. Given exogenous fitness, ω_0, the expected fitness values are

$$\begin{aligned}
\mathbb{E}[\Pi_1(D;D)] &= \omega_0 + \sum_{n=1}^{\infty} \delta^{n-1} \left\{1 - \psi^{2n-2}\tilde{x} + b\psi^{2n-1}\tilde{x}\right\} \\
&= \omega_0 + \frac{1}{1-\delta} + \frac{\tilde{x}\,(b\psi - 1)}{1-\delta\psi^2} \\
\mathbb{E}[\Pi_2(D;D)] &= \omega_0 + \sum_{n=1}^{\infty} \delta^{n-1} \left\{1 - \psi^{2n-1}\tilde{x} + b\psi^{2n-2}\tilde{x}\right\} \\
&= \omega_0 + \frac{1}{1-\delta} + \frac{\tilde{x}\,(b - \psi)}{1-\delta\psi^2}.
\end{aligned} \tag{16}$$

We assume that individuals are randomly assigned to the role of first and second movers with equal probability. Thus, the expected fitness of having the resident de-escalating strategy is

$$\begin{aligned}
W(D) &= (1/2)\left\{\mathbb{E}[\Pi_1(D;D)] + \mathbb{E}[\Pi_2(D;D)]\right\} \\
&= \omega_0 + \frac{1}{1-\delta} + \frac{\tilde{x}(b-1)(1+\psi)}{2\,(1-\delta\psi^2)}.
\end{aligned} \tag{17}$$

Consider a rare mutant de-escalator, D', with a strategy characterised by initial transfer $\tilde{z}$ and de-

escalation ρ. The expected fitness values for such a rare mutant, when in the roles of first mover and second mover respectively, are

$$\begin{aligned}
\mathbb{E}[\Pi_1(D';D)] &= \omega_0 + \sum_{n=1}^{\infty} \delta^{n-1} \left\{1 - \rho^{n-1}\psi^{n-1}\tilde{z} + b\rho^{n-1}\psi^{n}\tilde{z}\right\} \\
&= \omega_0 + \frac{1}{1-\delta} + \frac{\tilde{z}\,(b\psi - 1)}{1 - \delta\rho\psi} \\
\mathbb{E}[\Pi_2(D';D)] &= \omega_0 + \sum_{n=1}^{\infty} \delta^{n-1} \left\{1 - \rho^{n}\psi^{n-1}\tilde{x} + b\rho^{n-1}\psi^{n-1}\tilde{x}\right\} \\
&= \omega_0 + \frac{1}{1-\delta} + \frac{\tilde{x}\,(b - \rho)}{1 - \delta\rho\psi}.
\end{aligned} \tag{18}$$

The expected fitness of the rare mutant is

$$\begin{aligned}
W(D') &= (1/2)\left\{\mathbb{E}[\Pi_1(D';D)] + \mathbb{E}[\Pi_2(D';D)]\right\} \\
&= \omega_0 + \frac{1}{1-\delta} + \frac{\tilde{z}(b\psi - 1) + \tilde{x}(b - \rho)}{2\,(1 - \delta\rho\psi)}.
\end{aligned} \tag{19}$$

To check for an interior equilibrium that is both resistant to local invasion and convergent stable[1], the first-order necessary conditions are the following,

$$\begin{aligned}
\left.\frac{\partial W(D')}{\partial \tilde{z}}\right|_{\substack{\tilde{z}=\tilde{x} \\ \rho=\psi}} &= \frac{b\psi - 1}{2(1 - \delta\psi^2)} = 0 \\
\left.\frac{\partial W(D')}{\partial \tilde{\rho}}\right|_{\substack{\tilde{z}=\tilde{x} \\ \rho=\psi}} &= \frac{\tilde{x}\,(\delta b\psi^2 + \delta(b-1)\psi - 1)}{2(1 - \delta\psi^2)^2} = 0.
\end{aligned} \tag{20}$$

Satisfying both conditions in (20) requires that $\delta = 1$, which is false by assumption, and so we have no candidate equilibrium in the interior.

1.1.6 A graph of local mutations for a full lattice of de-escalating monomorphisms

To make progress, we adopt a graphical approach. Specifically, the analysis of monomorphisms for continuous strategy spaces centres on two basic ideas[1,2]. First, we can think of the resident strategy

as a candidate equilibrium strategy. If nearby strategies have strictly lower fitness, the resident strategy is resistant to invasion by local mutant strategies. Second, we can think of the resident strategy as a strategy near a candidate equilibrium. Loosely speaking, if mutations nearer to the candidate equilibrium than the resident strategy have strictly higher fitness than the resident type, the candidate equilibrium is convergent stable. In multidimensional continuous strategy spaces like those we consider, convergence stability comes in different forms. These different forms vary in terms of the restrictions they do or do not place on the structure of mutations, specifically restrictions related to how mutations correlate across the dimensions of the strategy space[3].

In our simulations (§ 2), mutations are uncorrelated across dimensions. We do not, however, impose such restrictions here. Instead, we develop a graphical method that creates a lattice over the entire strategy space. For any given point in the lattice, we treat the point as a monomorphism and plot it in black. We then create another small lattice for points around this resident strategy. These points represent rare mutations. For a given mutation, if the resident strategy has a strictly higher fitness, we plot the mutational point in red (i.e. a "red light" for the mutation). If the resident strategy has strictly lower fitness than the mutation, we plot the mutational point in green (i.e. a "green light" for the mutation). Whether red or green, colour intensity additionally signifies the magnitude of the difference in fitness values. If the two strategies have the same fitness, we plot the mutational point in white.

Specifically, consider a lattice of resident strategies, $D = (\tilde{x}, \psi) \in \{0, 0.1, \ldots, 1\}^2$. For a given point on the lattice, further consider a lattice of local mutations, $D' = (\tilde{x} + \epsilon_1, \psi + \epsilon_2)$, where $(\epsilon_1, \epsilon_2) \in \{-0.03, -0.02, -0.01, 0.01, 0.02, 0.03\}^2$. We treat $(\tilde{x}, \psi)$ as the resident strategy and $(\tilde{x} + \epsilon_1, \psi + \epsilon_2)$ as a rare mutant[a]. At the point $(\tilde{x} + \epsilon_1, \psi + \epsilon_2)$, we plot $W(D) - W(D')$. If $W(D) - W(D') > 0$, we plot the point in red. If $W(D) - W(D') < 0$, we plot the point in green.

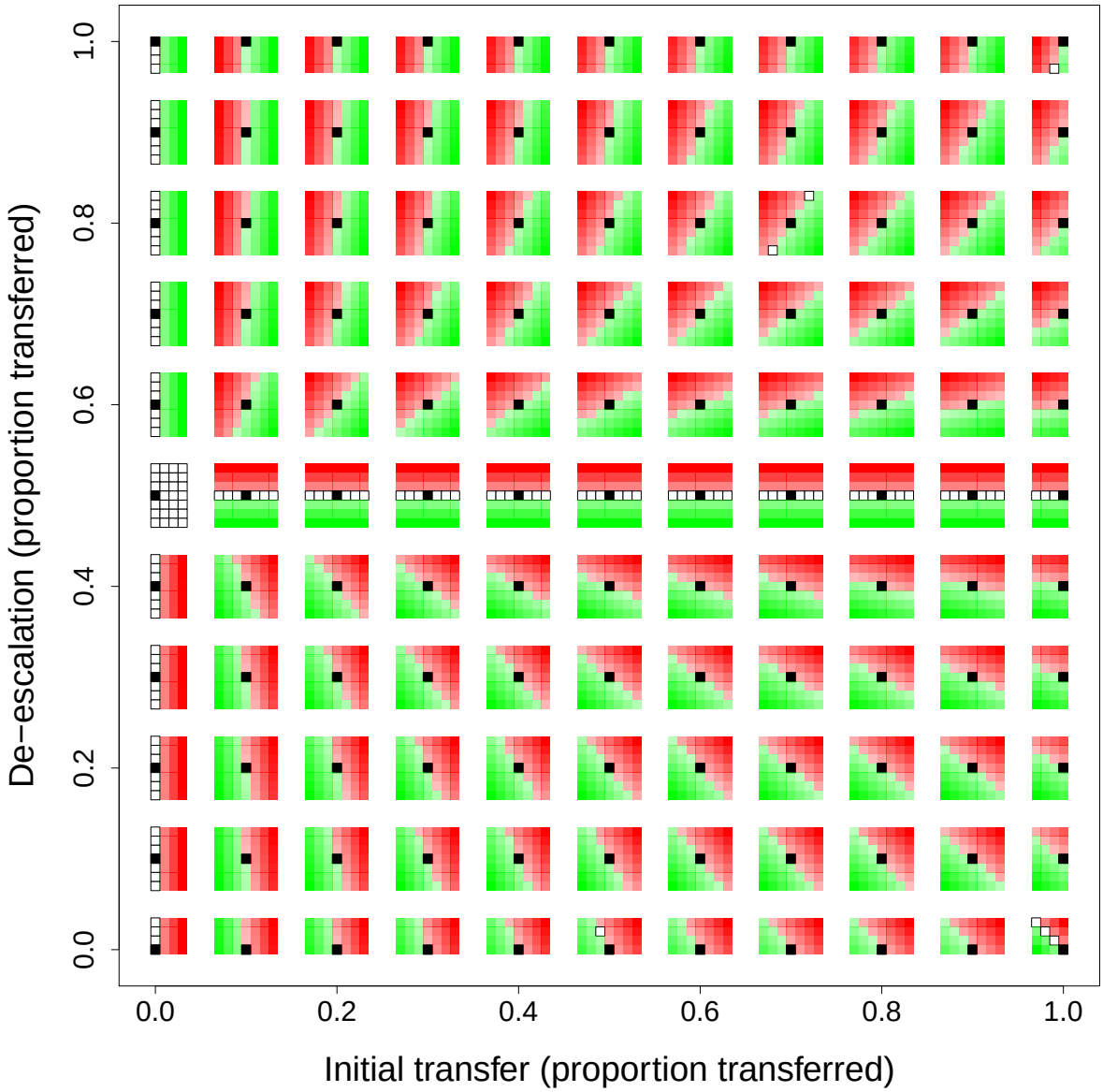
[a]Please note we restrict attention to strategies in the unit square, which means we simply ignore mutations that would take us outside this region

Whether red or green, colour intensity represents the magnitude of the difference in fitness values between the resident strategy and the rare mutant. The point in the ***local*** mutational field with the fitness difference of largest magnitude has the most intense colour. In this sense, we examine local mutations relative to a given resident strategy in a population that is, aside from the rare mutant, monomorphic. This local analysis for one point on the lattice of monomorphisms is not directly linked to the local analysis for any other point on the lattice of monomorphisms.

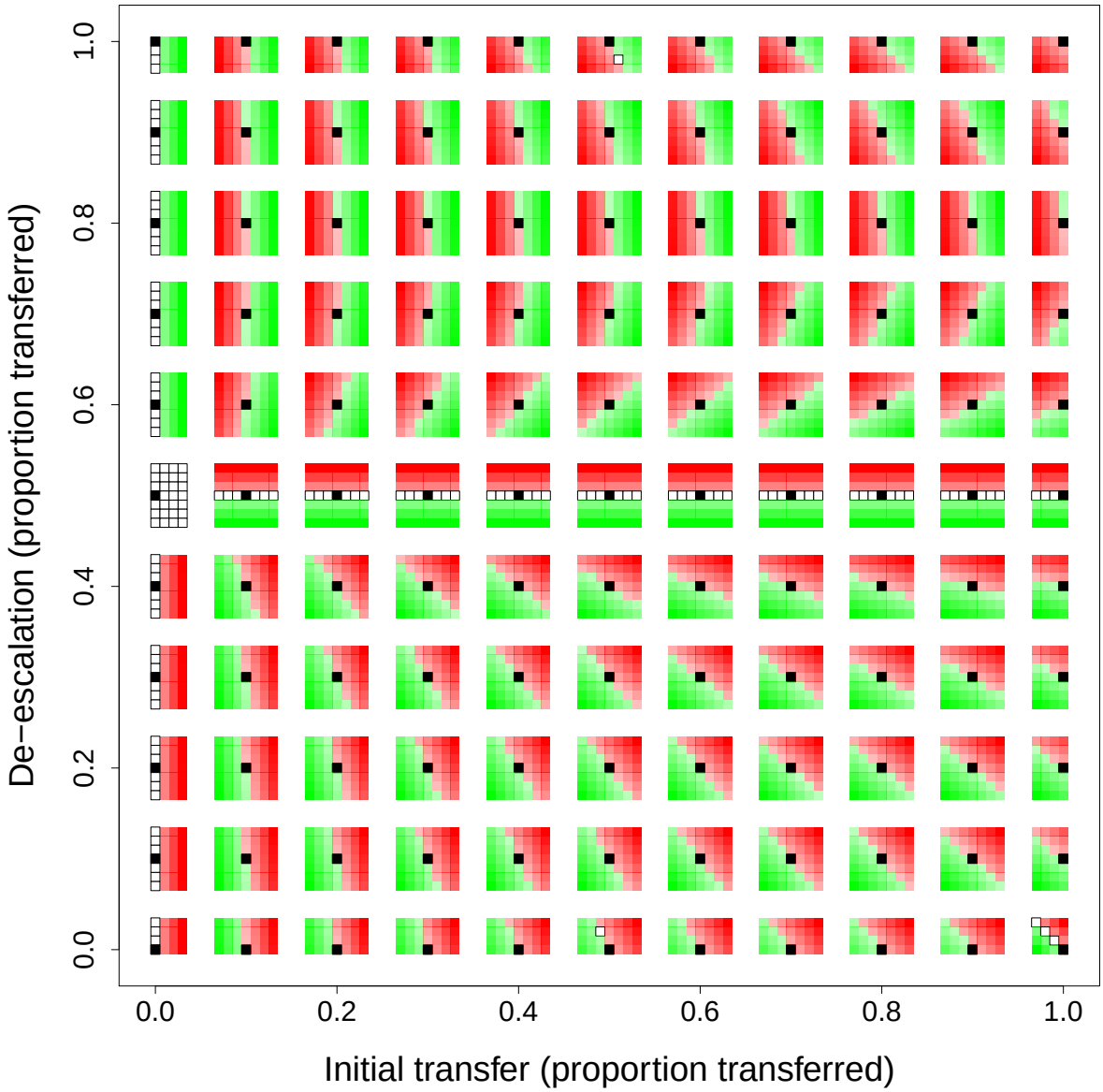
Supplementary Figures 1-4 show graphs of this sort based on four different values of δ from $\{0.25, 0.5, 0.9, 0.99\}$. The graphs show the following.

- If both the resident strategy and the rare mutant begin an interaction by transferring nothing, all forms of de-escalation are neutral with respect to each other.
- If the resident strategy is such that $\psi = 1/b$, all local mutations in which $\rho = \psi$ and $\tilde{z} \neq \tilde{x}$ are neutral with respect to the resident strategy.
- If the resident strategy involves a relatively small de-escalation value (i.e. strong de-escalation), monomorphisms are susceptible to invasion by rare mutants with smaller initial transfers and smaller de-escalation values (i.e. stronger de-escalation). Moreover, resident strategies involving an initial transfer of zero are resistant to invasion by rare local mutants with positive initial transfer values.
- If δ is relatively low, and if the resident strategy involves a relatively large de-escalation value (i.e. weak de-escalation), monomorphisms are susceptible to invasion by rare mutants with larger initial transfers and smaller de-escalation values (i.e. stronger de-escalation).
- If δ is relatively high, and if the resident strategy involves a relatively large de-escalation value (i.e. weak de-escalation), monomorphisms are susceptible to invasion by rare mutants with larger initial transfers and larger de-escalation values (i.e. weaker de-escalation).

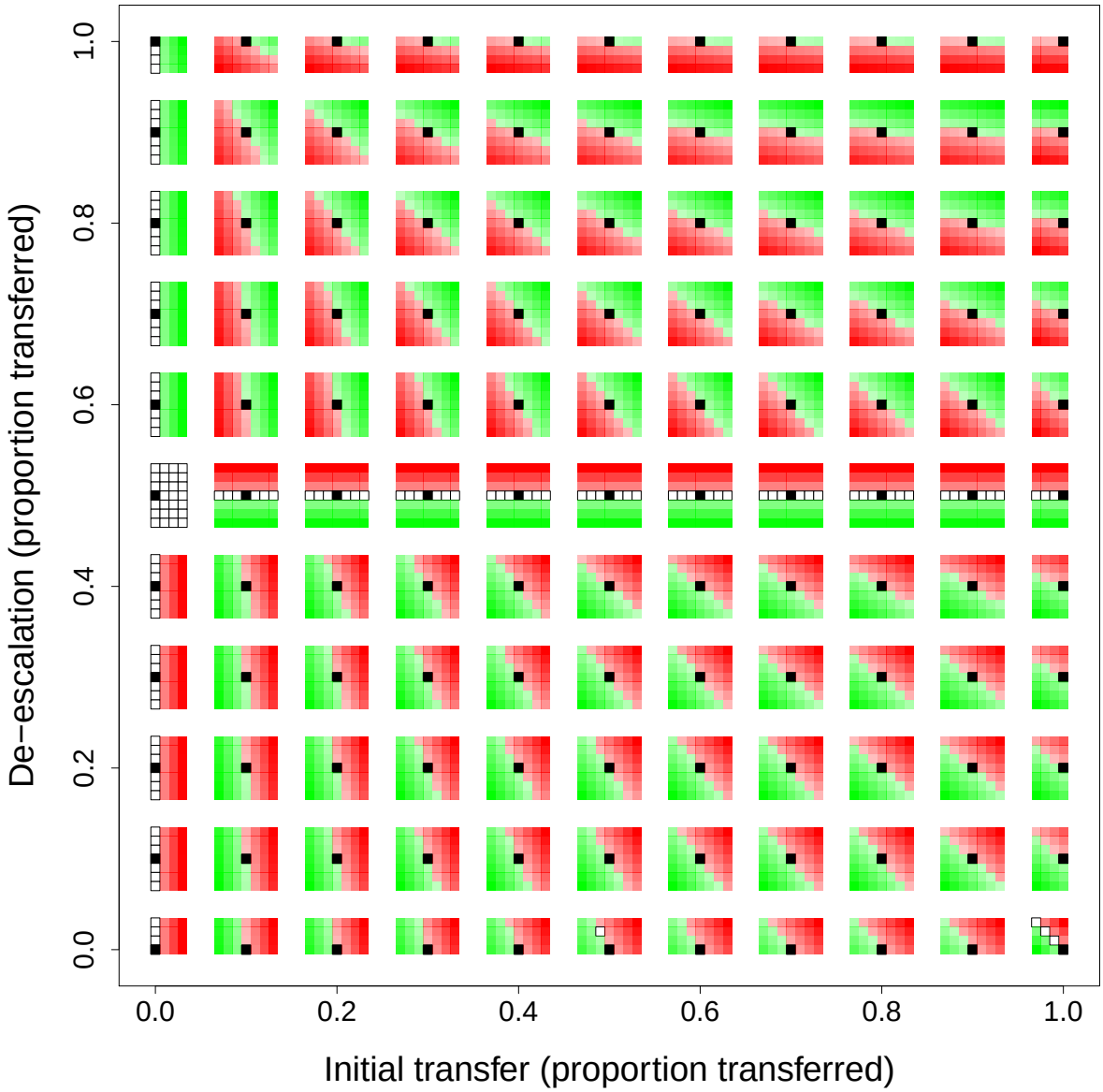
In sum, if the resident strategy involves relatively strong de-escalation, the population should converge on increasingly selfish strategies. In the limit, initial transfer values approach zero, and all forms of de-escalation become equivalent. At this point, we expect drift to be relatively important with respect to the subsequent evolution of de-escalation values. If the resident strategy involves relatively weak de-escalation, dynamics should depend on δ. If δ is relatively small, stronger de-escalation should evolve until the population enters the region of strategy space in which selection uniformly favours increasingly selfish strategies. If δ is sufficiently large, increasingly weak forms of de-escalation should evolve until the population converges on perfect reciprocity, which is the boundary between de-escalating reciprocity and escalating reciprocity.



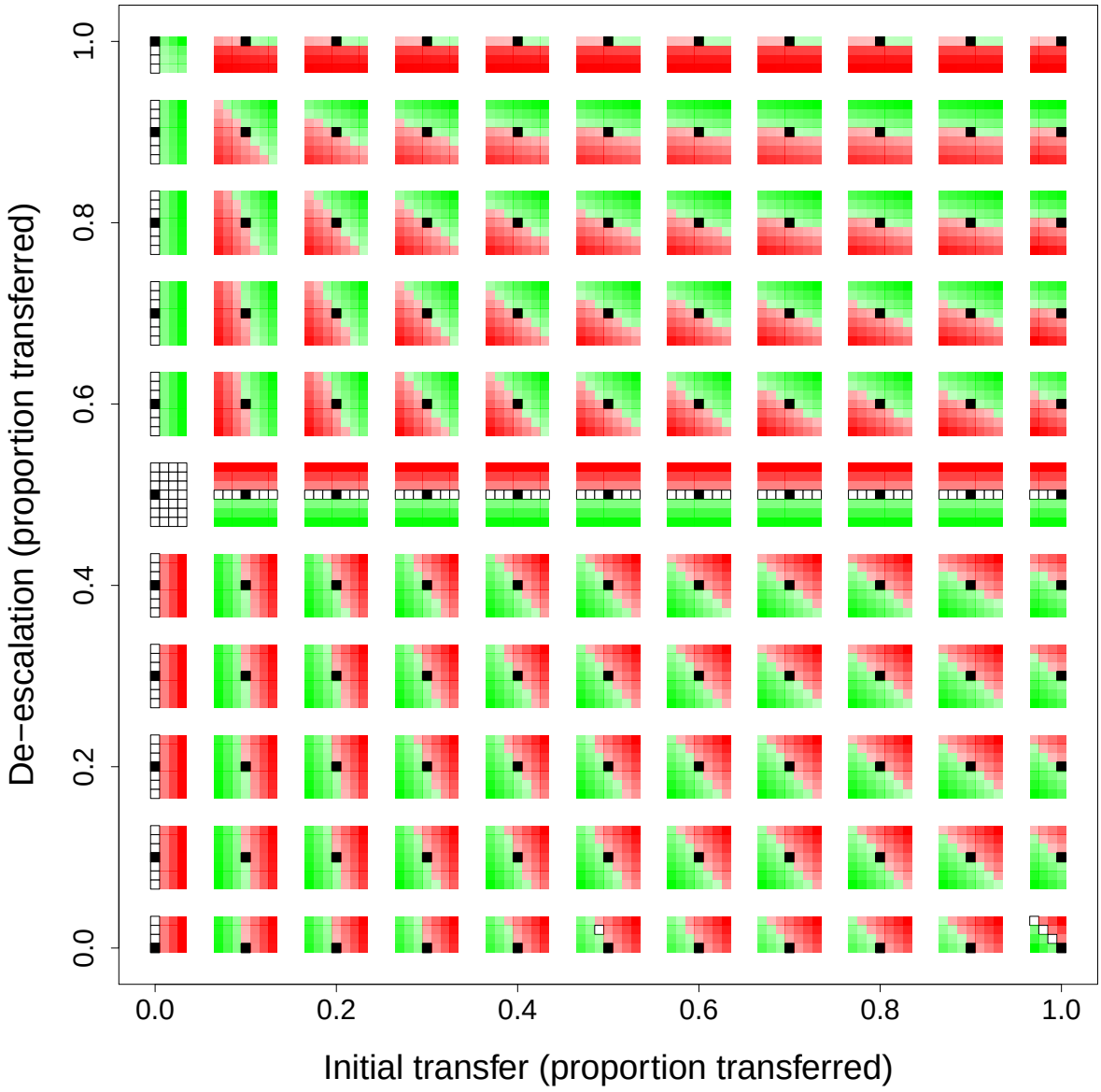
Supplementary Figure 1 | De-escalating resident strategies with local mutations, $\delta = 0.25$. Each black point is a resident strategy. The surrounding cloud of points shows the fitness effects of rare local mutations. Green signifies that the rare local mutation has higher fitness than the resident strategy, while red indicates the mutation has lower fitness than the resident strategy. Colour intensity represents the magnitude of the fitness difference. A white point within a cloud of local mutations means the resident strategy and the rare mutant have the same fitness. Initial transfer values are expressed as the proportion of the endowment transferred to one's partner, and so higher values indicate more cooperation. De-escalation values are expressed as the proportion of the endowment transferred as a function of the partner's most recent transfer. Higher values represent weaker de-escalation and thus more cooperation. See § 1.1.6 for a detailed explanation of the graph.



Supplementary Figure 2 | De-escalating resident strategies with local mutations, $\delta = 0.5$. Each black point is a resident strategy. The surrounding cloud of points shows the fitness effects of rare local mutations. Green signifies that the rare local mutation has higher fitness than the resident strategy, while red indicates the mutation has lower fitness than the resident strategy. Colour intensity represents the magnitude of the fitness difference. A white point within a cloud of local mutations means the resident strategy and the rare mutant have the same fitness. Initial transfer values are expressed as the proportion of the endowment transferred to one's partner, and so higher values indicate more cooperation. De-escalation values are expressed as the proportion of the endowment transferred as a function of the partner's most recent transfer. Higher values represent weaker de-escalation and thus more cooperation. See § 1.1.6 for a detailed explanation of the graph.



Supplementary Figure 3 | De-escalating resident strategies with local mutations, $\delta = 0.9$. Each black point is a resident strategy. The surrounding cloud of points shows the fitness effects of rare local mutations. Green signifies that the rare local mutation has higher fitness than the resident strategy, while red indicates the mutation has lower fitness than the resident strategy. Colour intensity represents the magnitude of the fitness difference. A white point within a cloud of local mutations means the resident strategy and the rare mutant have the same fitness. Initial transfer values are expressed as the proportion of the endowment transferred to one's partner, and so higher values indicate more cooperation. De-escalation values are expressed as the proportion of the endowment transferred as a function of the partner's most recent transfer. Higher values represent weaker de-escalation and thus more cooperation. See § 1.1.6 for a detailed explanation of the graph.



Supplementary Figure 4 | De-escalating resident strategies with local mutations, $\delta = 0.99$. Each black point is a resident strategy. The surrounding cloud of points shows the fitness effects of rare local mutations. Green signifies that the rare local mutation has higher fitness than the resident strategy, while red indicates the mutation has lower fitness than the resident strategy. Colour intensity represents the magnitude of the fitness difference. A white point within a cloud of local mutations means the resident strategy and the rare mutant have the same fitness. Initial transfer values are expressed as the proportion of the endowment transferred to one's partner, and so higher values indicate more cooperation. De-escalation values are expressed as the proportion of the endowment transferred as a function of the partner's most recent transfer. Higher values represent weaker de-escalation and thus more cooperation. See § 1.1.6 for a detailed explanation of the graph.

1.1.7 An analysis of escalating monomorphisms

Now consider a monomorphic population of escalators, E, characterised by initial transfers, $\tilde{x}$, and a degree of escalation, $\alpha \in [0,1]$. $\Pi_1(E;E)$ is a random variable with support $\mathbb{R}_+$ that denotes the fitness of an escalating first mover of the resident type with an escalating partner of the same type. $\Pi_2(E;E)$ is the analogous random variable for an escalating second mover of the resident type with a partner of the same type.

We can simplify the analysis greatly by working with a straightforward transformation of actions and associated strategies. Accordingly, let $r_n = 1 - x_n$ be the proportion of the first mover's endowment retained from interaction n. Similarly, $t_n = 1 - y_n$ is the proportion of the second mover's endowment retained from n. To specify strategies for escalators of the resident monomorphism, let $\overline{r} = 1 - \tilde{x}$ be the proportion of the endowment retained by such an individual, if first mover, in $n = 1$. Moreover, let $\xi = 1 - \alpha$ specify how the individual reduces, perhaps only weakly, what is retained with respect to her partner's most recent move. Altogether, strategies take the form,

$$\begin{aligned} n = 1, \quad & r_n = \overline{r} \\ & t_n = \xi \overline{r} \\ n \geq 2, \quad & r_n = \xi t_{n-1} \\ & t_n = \xi r_n. \end{aligned} \tag{21}$$

For a pair of this sort, the general solution for their choices is

$$\begin{aligned} n \geq 1, \quad & r_n = \xi^{2n-2} \overline{r} \\ & t_n = \xi^{2n-1} \overline{r}. \end{aligned} \tag{22}$$

By extension, the expected fitness values conditional on one's role as first and second mover are,

respectively,

$$\begin{aligned}\mathbb{E}[\Pi_1(E;E)] &= \omega_0 + \sum_{n=1}^{\infty} \delta^{n-1} \left\{\xi^{2n-2}\overline{r} + b(1-\xi^{2n-1}\overline{r})\right\} \\ &= \omega_0 + \frac{b}{1-\delta} + \frac{\overline{r}(1-b\xi)}{1-\delta\xi^2} \\ \mathbb{E}[\Pi_2(E;E)] &= \omega_0 + \sum_{n=1}^{\infty} \delta^{n-1} \left\{\xi^{2n-1}\overline{r} + b(1-\xi^{2n-2}\overline{r})\right\} \\ &= \omega_0 + \frac{b}{1-\delta} + \frac{\overline{r}(\xi-b)}{1-\delta\xi^2}.\end{aligned} \tag{23}$$

The expected fitness of an escalator with the resident strategy is thus

$$\begin{aligned}W(E) &= (1/2)\left\{\mathbb{E}[\Pi_1(E;E)] + \mathbb{E}[\Pi_2(E;E)]\right\} \\ &= \omega_0 + \frac{b}{1-\delta} + \frac{\overline{r}(1-b)(1+\xi)}{2(1-\delta\xi^2)}.\end{aligned} \tag{24}$$

Now consider a rare mutant, E', who retains $\overline{w}$ as first mover and who escalates by retaining some proportion, η, of her partner's most recent move. The expected fitness values of such a rare mutant are

$$\begin{aligned}\mathbb{E}[\Pi_1(E';E)] &= \omega_0 + \sum_{n=1}^{\infty} \delta^{n-1} \left\{\eta^{n-1}\xi^{n-1}\overline{w} + b(1-\eta^{n-1}\xi^{n}\overline{w})\right\} \\ &= \omega_0 + \frac{b}{1-\delta} + \frac{\overline{w}(1-b\xi)}{1-\delta\eta\xi} \\ \mathbb{E}[\Pi_2(E';E)] &= \omega_0 + \sum_{n=1}^{\infty} \delta^{n-1} \left\{\eta^{n}\xi^{n-1}\overline{r} + b(1-\eta^{n-1}\xi^{n-1}\overline{r})\right\} \\ &= \omega_0 + \frac{b}{1-\delta} + \frac{\overline{r}(\eta-b)}{1-\delta\eta\xi}.\end{aligned} \tag{25}$$

The expected fitness of the rare mutant is

$$\begin{aligned}W(E') &= (1/2)\left\{\mathbb{E}[\Pi_1(E';E)] + \mathbb{E}[\Pi_2(E';E)]\right\} \\ &= \omega_0 + \frac{b}{1-\delta} + \frac{\overline{w}(1-b\xi) + \overline{r}(\eta-b)}{2(1-\delta\eta\xi)}.\end{aligned} \tag{26}$$

To check for an interior equilibrium that is both resistant to local invasion and convergent

stable[1], the first-order necessary conditions are the following,

$$\begin{aligned}\left.\frac{\partial W(E')}{\partial \tilde{w}}\right|_{\substack{\tilde{w}=\tilde{r}\\ \eta=\xi}} &= \frac{1-b\xi}{2(1-\delta\xi^2)} = 0\\ \left.\frac{\partial W(E')}{\partial \tilde{\eta}}\right|_{\substack{\tilde{x}=\tilde{r}\\ \eta=\xi}} &= \frac{-\tilde{r}\left(\delta b\xi^2 - \delta(1-b)\xi - 1\right)}{2(1-\delta\xi^2)^2} = 0.\end{aligned} \tag{27}$$

Satisfying both conditions in (27) requires that $\delta = 1$, which is false by assumption, and so we have no candidate equilibrium in the interior.

1.1.8 A graph of local mutations for a full lattice of escalating monomorphisms

We use a graphical method precisely analogous to that explained in § 1.1.6. Consider a lattice of resident strategies, $E = (\tilde{r}, \xi) \in \{0, 0.1, \ldots, 1\}^2$. For given point on this lattice, further consider a lattice of local mutations, $E' = (\tilde{r}+\epsilon_1, \xi+\epsilon_2)$, where $(\epsilon_1, \epsilon_2) \in \{-0.03, -0.02, -0.01, 0.01, 0.02, 0.03\}^2$. We treat $(\tilde{r}, \xi)$ as the resident strategy and $(\tilde{r} + \epsilon_1, \xi + \epsilon_2)$ as a rare mutant[b]. At the point $(\tilde{r} + \epsilon_1, \xi + \epsilon_2)$, we plot $W(E) - W(E')$. If $W(E) - W(E') > 0$, we plot the point in red. If $W(E) - W(E') < 0$, we plot the point in green. Whether red or green, colour intensity represents the magnitude of the difference in fitness values between the resident strategy and the rare mutant. The point in the ***local*** mutational field with the fitness difference of largest magnitude has the most intense colour. In this sense, we examine local mutations relative to a given resident strategy in a population that is, aside from the rare mutant, monomorphic. This local analysis for one point on the lattice of monomorphisms is not directly linked to the local analysis for any other point on the lattice of monomorphisms.

For our analysis of escalating strategies, we make one addition to this graphical approach. Under certain restrictions, there exists an interior degree of escalation that strictly dominates local

[b]Please note we restrict attention to strategies in the unit square, which means we simply ignore mutations that would take us outside this region

deviations if we limit attention to deviations involving the degree of escalation. Specifically, assume the mutant strategy only involves the degree of escalation, and so $\overline{w} = \overline{r}$. We restrict attention to $\overline{r} > 0$ because all forms of escalation are payoff-equivalent if $\overline{r} = 0$. Define the function $f : \mathbb{R} \to \mathbb{R}$ such that $f(\xi) = 1 + \delta\xi(1 - b(1 + \xi))$. If $\overline{r} > 0$, $W(E) > W(E')$ if $(\eta - \xi)f(\xi) < 0$. If $\delta > 0$ and $b > 1$, $\forall \xi \in \mathbb{R}$, f is a concave function with one negative and one positive root. Denote the positive root as

$$\hat{\xi}_1 = \frac{\delta(1-b) + \sqrt{\delta^2(1-b)^2 + 4b\delta}}{2b\delta}. \tag{28}$$

Because f is concave, $\forall \xi \in [0, \hat{\xi}_1)$, $f(\xi) > 0$, while $\forall \xi > \hat{\xi}_1$, $f(\xi) < 0$. In our model, ξ only has meaning when $\xi \in [0, 1]$, and we will restrict attention to this interval. In particular, if $(1 + \delta)/(2\delta) < b$, $\hat{\xi}_1 < 1$. If $(1 + \delta)/(2\delta) > b$, $\hat{\xi}_1 > 1$.

Thus, if $(1 + \delta)/(2\delta) > b$, $\hat{\xi}_1 > 1$ and, $\forall \xi \in [0, 1]$, $f(\xi) > 0$. Consequently, $W(E) > W(E')$ if $\xi > \eta$, which is equivalent to saying the resident strategy resists invasion by the rare mutant if the resident strategy retains more (i.e. escalates in a weaker fashion) than the mutant. In this case, no interior degree of escalation resists invasion by local mutant strategies.

Alternatively, if $(1 + \delta)/(2\delta) < b$, $\hat{\xi}_1 \in (0, 1)$. In this case, $W(E) > W(E')$ if $\eta > \hat{\xi}_1$. Moreover, $W(E) > W(E')$ if $\hat{\xi}_1 > \eta$. This means that, if we restrict attention to rare mutants who deviate only in terms of escalation, $\hat{\xi}_1$ strictly dominates all local deviations and is thus resistant to invasion. When $\hat{\xi}_1 \in (0, 1)$, we add it to our graphs as a heavy dashed line in black. The idea is simply to see if local mutation fields suggest convergence to a region of strategy space that includes $\hat{\xi}_1$. When interpreting these graphs, however, please bear certain caveats in mind. First, the colours show the fitness effects of local mutations relative to resident strategies on the lattice. If $\hat{\xi}_1$ falls near a degree of escalation that is on the lattice, but $\hat{\xi}_1$ is itself not on the lattice, the colours refer to deviations from the point on the lattice, not deviations from $\hat{\xi}_1$. Second, the two types of analysis may suggest mutually congruent fitness effects, but they are not the same. The

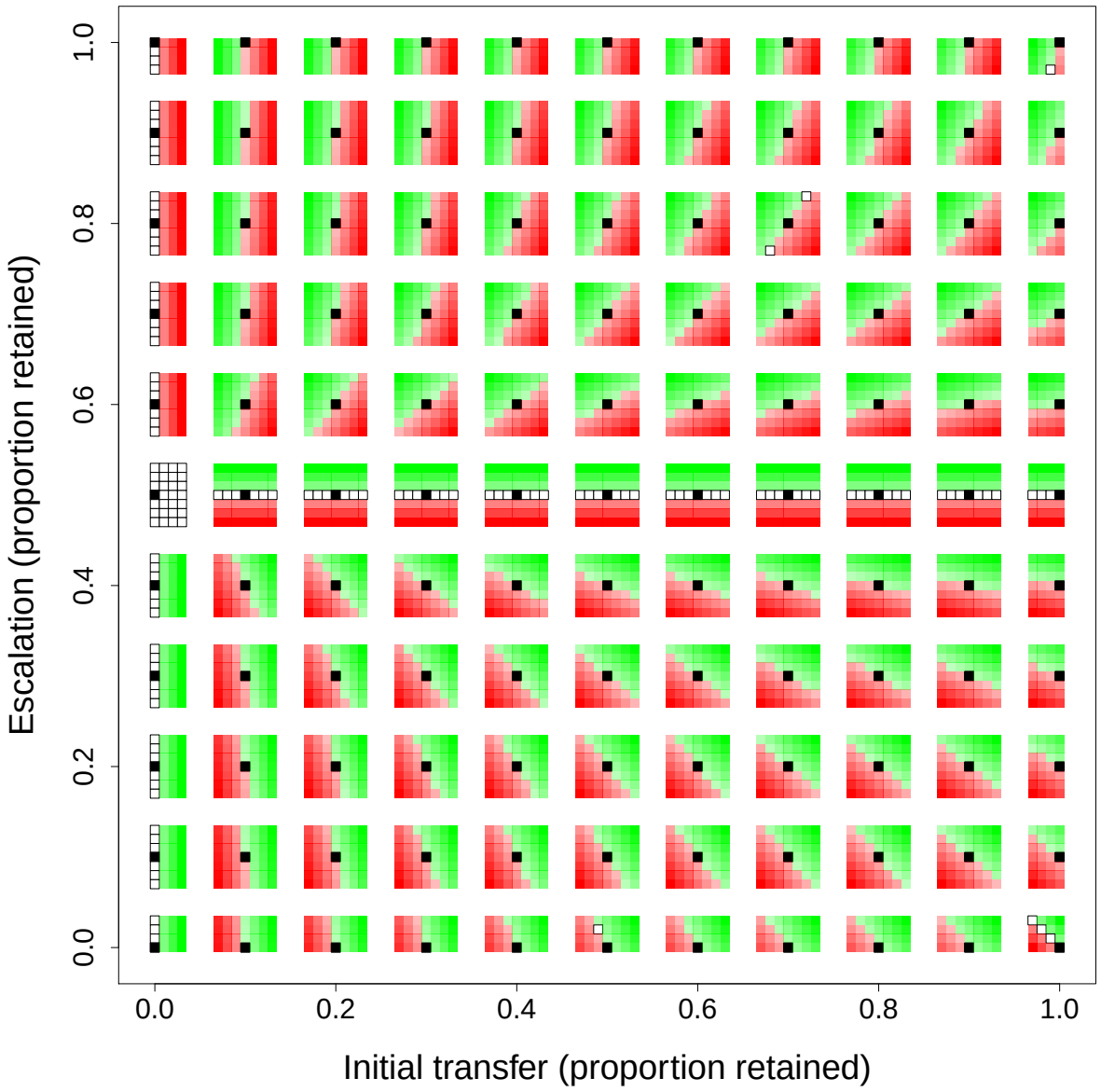
colours show the effects of deviations in any arbitrary direction from a resident monomorphism. In contrast, identifying $\hat{\xi}_1$ rests on an approach that assumes $\overline{w} = \overline{r} > 0$ and additionally posits mutations in one dimension of strategy space only.

Supplementary Figures 5-8 show graphs based on four different values of δ from the set $\{0.25, 0.5, 0.9, 0.99\}$. The graphs show the following.

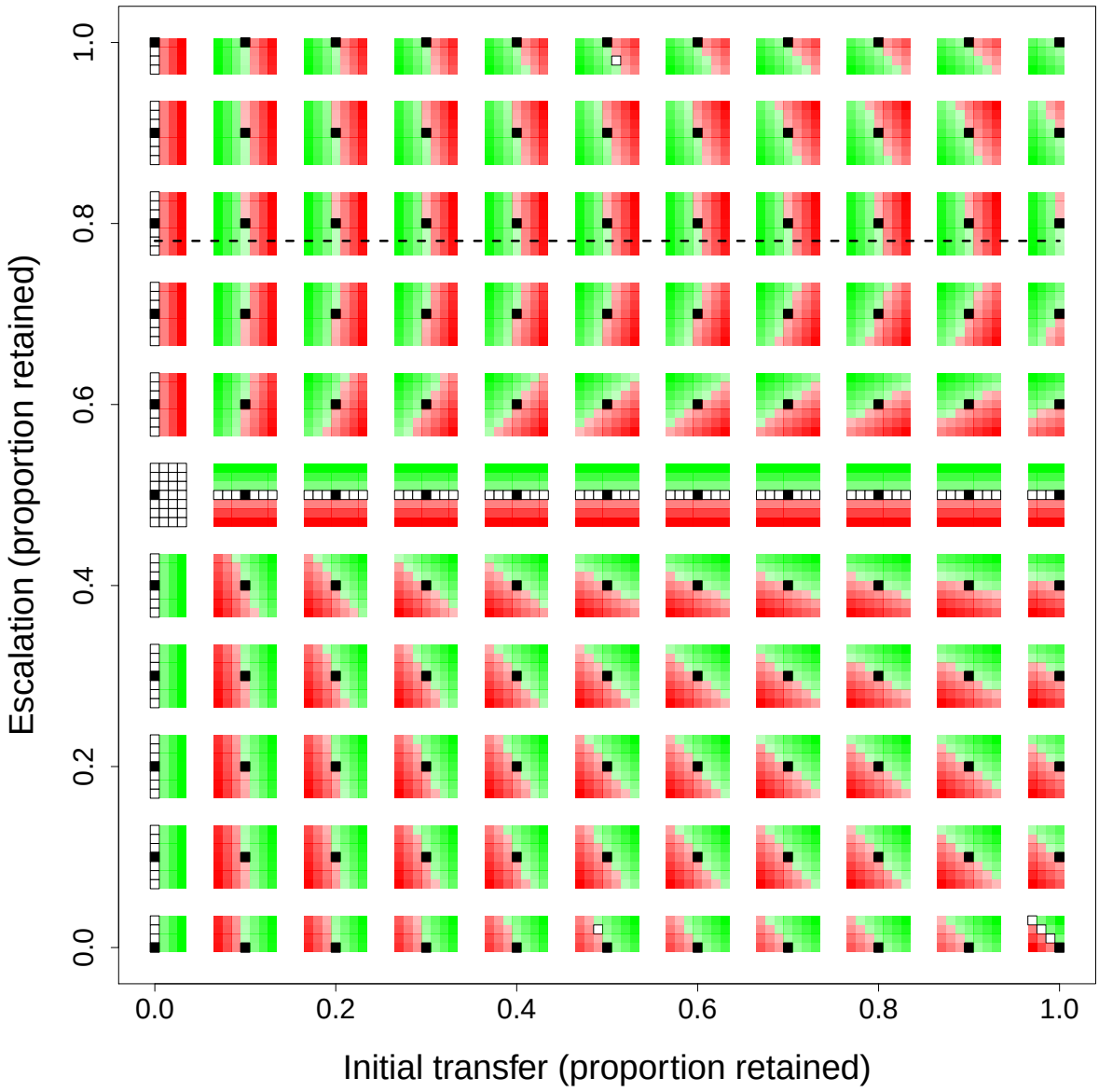
- If both the resident strategy and the rare mutant begin an interaction by transferring everything (i.e. retaining zero), all forms of escalation are neutral with respect to each other.
- If the resident strategy is such that $\xi = 1/b$, all local mutations in which $\eta = \xi$ and $\tilde{w} \neq \tilde{r}$ are neutral with respect to the resident strategy.
- If the resident strategy involves a relatively small escalation value (i.e. strong escalation because strategy retains little), monomorphisms are susceptible to invasion by rare mutants with larger initial transfers (i.e. retain more) and larger escalation values (i.e. weaker escalation because strategies retain more). Moreover, resident strategies involving an initial transfer value of zero (i.e. entire endowment transferred because zero retained) are susceptible to invasion by rare local mutants with positive initial transfer values (i.e. strategies that retain some positive proportion of endowment).
- If δ is relatively low, and if the resident strategy involves a relatively large escalation value (i.e. weak escalation because strategy retains much), resident strategies are resistant to invasion by rare mutants with larger initial transfer values (i.e. retain more) and smaller escalation values (i.e. stronger escalation because strategy retains less).
- If δ is relatively high, and if the resident strategy involves a relatively large escalation value (i.e. weak escalation), resident strategies are susceptible to invasion by rare mutants with

smaller initial transfer values (i.e. retain less) and smaller escalation values (i.e. stronger escalation).

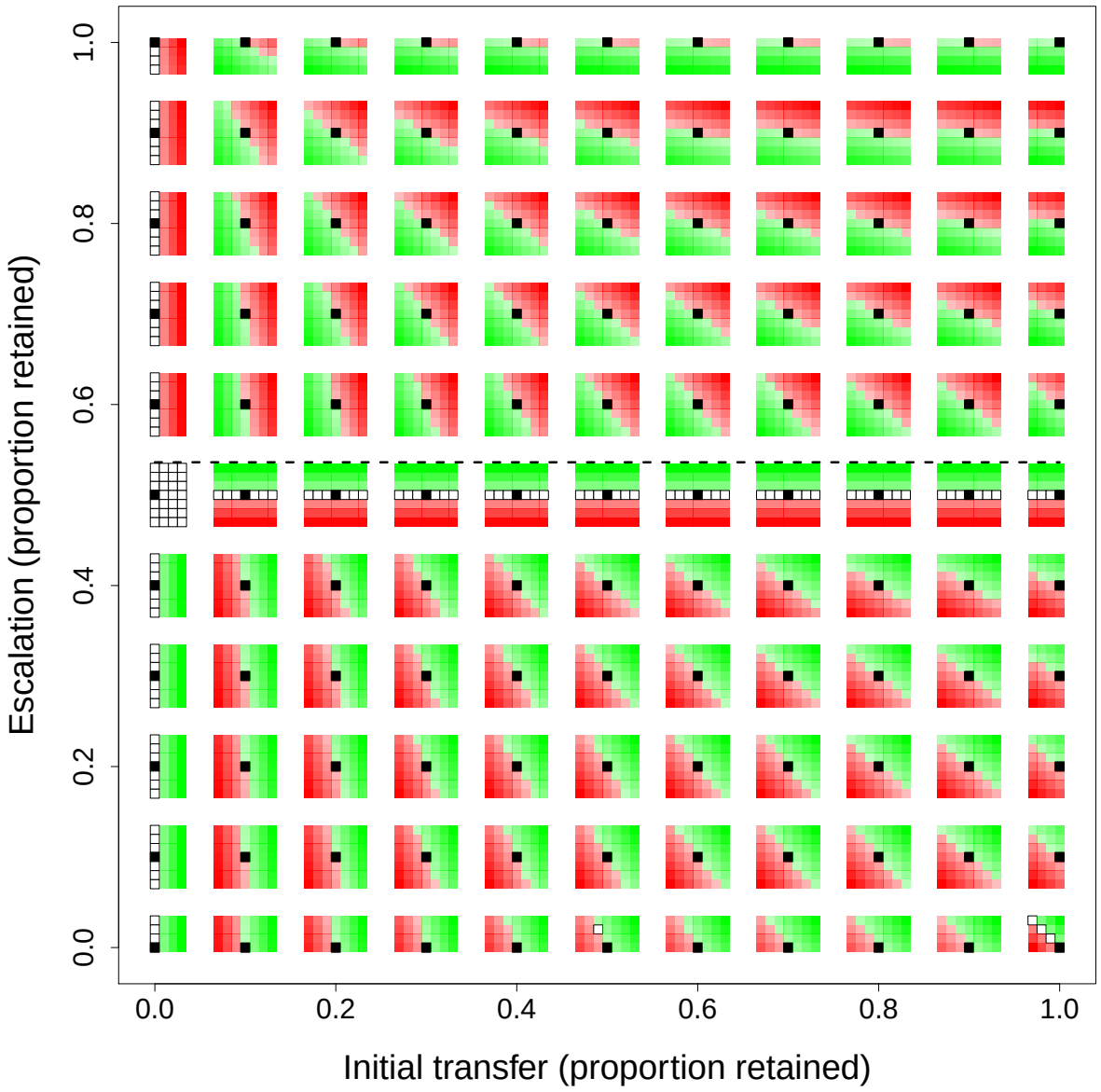
In sum, if the resident strategy is extremely cooperative in the sense that it escalates strongly, the population should move toward strategies that are less cooperative in that they retain more for the initial move and escalate less strongly. If the resident strategy involves relatively weak escalation (i.e. large escalation values), the population should move toward smaller initial transfer values (i.e. less retained), and the dynamics of escalation should depend on δ. If δ is relatively small, weaker escalation should evolve until the population converges on perfect reciprocity, which is the boundary between de-escalating reciprocity and escalating reciprocity. Based on the earlier analysis of de-escalating monomorphisms (§ 1.1.6), we expect the population to continue moving toward increasingly selfish strategies. If δ is relatively large, the population should converge on an interior degree of escalation.



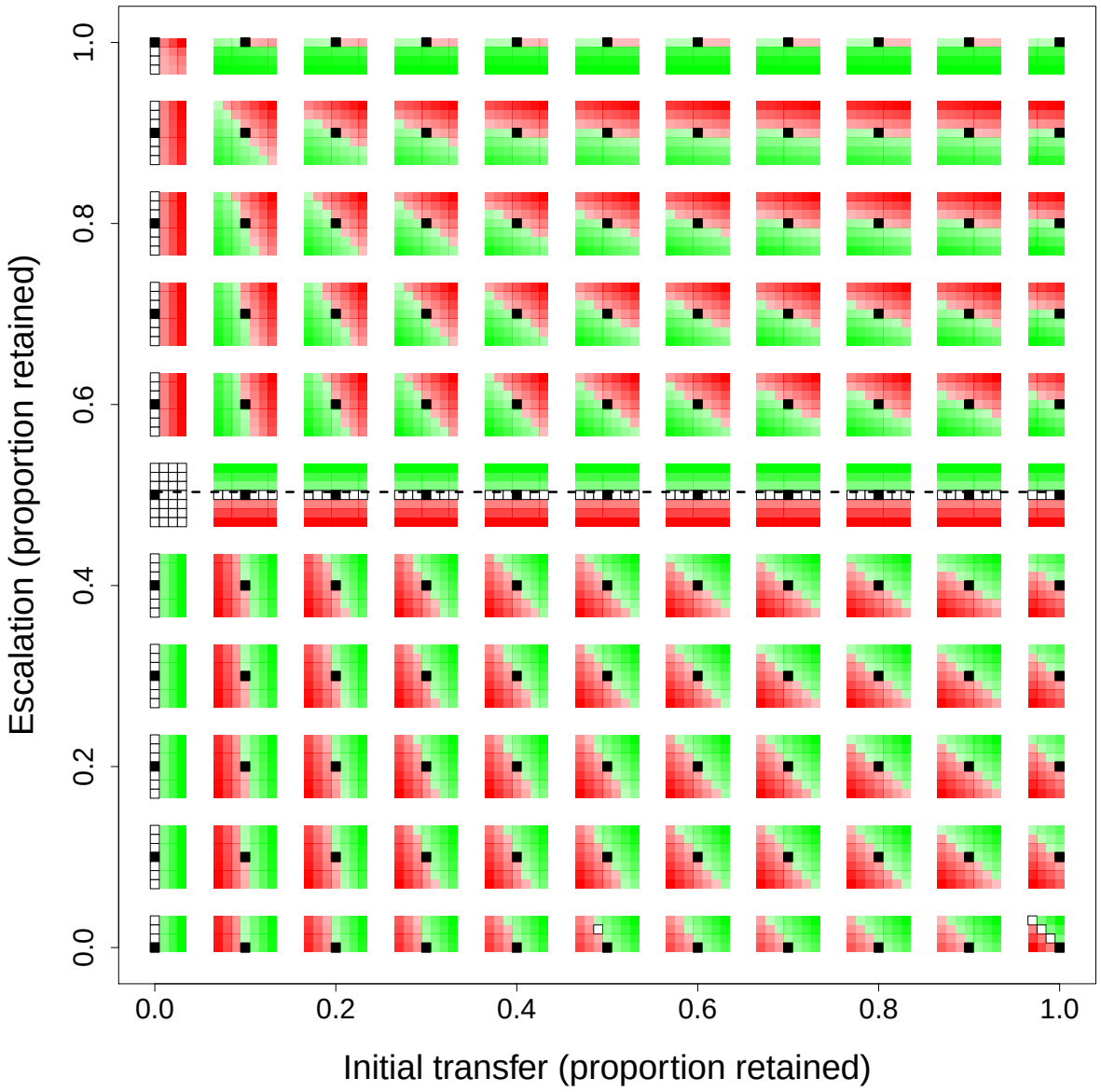
Supplementary Figure 5 | Escalating resident strategies with local mutations, $\delta = 0.25$. Each black point is a resident strategy. The surrounding cloud of points shows the fitness effects of rare local mutations. Green signifies that the rare local mutation has higher fitness than the resident strategy, while red indicates the mutation has lower fitness than the resident strategy. Colour intensity represents the magnitude of the fitness difference. A white point within a cloud of local mutations means the resident strategy and the rare mutant have the same fitness. Initial transfer values are expressed as the proportion of the endowment retained, and so higher values indicate less cooperation. Escalation values are expressed as the proportion of the endowment retained as a function of the proportion the partner most recently retained. Higher values represent weaker escalation and thus less cooperation. See § 1.1.8 for a detailed explanation of the graph.



Supplementary Figure 6 | Escalating resident strategies with local mutations, $\delta = 0.5$. Each black point is a resident strategy. The surrounding cloud of points shows the fitness effects of rare local mutations. Green signifies that the rare local mutation has higher fitness than the resident strategy, while red indicates the mutation has lower fitness than the resident strategy. Colour intensity represents the magnitude of the fitness difference. A white point within a cloud of local mutations means the resident strategy and the rare mutant have the same fitness. Initial transfer values are expressed as the proportion of the endowment retained, and so higher values indicate less cooperation. Escalation values are expressed as the proportion of the endowment retained as a function of the proportion the partner most recently retained. Higher values represent weaker escalation and thus less cooperation. See § 1.1.8 for a detailed explanation of the graph.



Supplementary Figure 7 | Escalating resident strategies with local mutations, $\delta = 0.9$. Each black point is a resident strategy. The surrounding cloud of points shows the fitness effects of rare local mutations. Green signifies that the rare local mutation has higher fitness than the resident strategy, while red indicates the mutation has lower fitness than the resident strategy. Colour intensity represents the magnitude of the fitness difference. A white point within a cloud of local mutations means the resident strategy and the rare mutant have the same fitness. Initial transfer values are expressed as the proportion of the endowment retained, and so higher values indicate less cooperation. Escalation values are expressed as the proportion of the endowment retained as a function of the proportion the partner most recently retained. Higher values represent weaker escalation and thus less cooperation. See § 1.1.8 for a detailed explanation of the graph.



Supplementary Figure 8 | Escalating resident strategies with local mutations, $\delta = 0.99$. Each black point is a resident strategy. The surrounding cloud of points shows the fitness effects of rare local mutations. Green signifies that the rare local mutation has higher fitness than the resident strategy, while red indicates the mutation has lower fitness than the resident strategy. Colour intensity represents the magnitude of the fitness difference. A white point within a cloud of local mutations means the resident strategy and the rare mutant have the same fitness. Initial transfer values are expressed as the proportion of the endowment retained, and so higher values indicate less cooperation. Escalation values are expressed as the proportion of the endowment retained as a function of the proportion the partner most recently retained. Higher values represent weaker escalation and thus less cooperation. See § 1.1.8 for a detailed explanation of the graph.

1.2 Three-dimensional strategies

The first mover's strategy involves three quantities. $\tilde{x}_i \in [0,1]$ is the initial transfer for $n = 1$, and $a_i, d_i \in [0,1]$ control the first mover's response function. The second mover's response function depends on $a_{i'}, d_{i'} \in [0,1]$. Strategies take the form,

$$\begin{aligned} n = 1, \quad & x_n = \tilde{x}_i \\ & y_n = a_{i'} + (d_{i'} - a_{i'})\tilde{x}_i \\ n \geq 2, \quad & x_n = a_i + (d_i - a_i)y_{n-1} \\ & y_n = a_{i'} + (d_{i'} - a_{i'})x_n. \end{aligned} \tag{29}$$

For a pair of this sort, the general solution for their choices is

$$\begin{aligned} n = 1, \quad & x_n = \tilde{x}_i \\ & y_n = a_{i'} + (d_{i'} - a_{i'})\tilde{x}_i \\ n \geq 2, \quad & x_n = (d_i - a_i)^{n-1}(d_{i'} - a_{i'})^{n-1}\tilde{x}_i + a_i \sum_{k=0}^{n-2}(d_i - a_i)^k(d_{i'} - a_{i'})^k \\ & \qquad + a_{i'} \sum_{k=0}^{n-2}(d_i - a_i)^{k+1}(d_{i'} - a_{i'})^k \\ & y_n = (d_i - a_i)^{n-1}(d_{i'} - a_{i'})^{n}\tilde{x}_i + a_i \sum_{k=0}^{n-2}(d_i - a_i)^k(d_{i'} - a_{i'})^{k+1} \\ & \qquad + a_{i'} \sum_{k=0}^{n-1}(d_i - a_i)^k(d_{i'} - a_{i'})^k. \end{aligned} \tag{30}$$

Below we will speak of unconditional types, reciprocators, and anti-reciprocators. Unconditional types have response functions with a slope of 0. Reciprocators have positively sloped response functions, and a perfect reciprocator has a response function with a slope of 1. Anti-reciprocators have negatively sloped response functions, and a perfect anti-reciprocator has a response function with a slope of -1. Finally, an escalator has a positively sloped response function that is above the 45-degree line and does not cross this line. A de-escalator has a positively sloped response function that is below the 45-degree line and does not cross this line. An ambiguous reciprocator

has a positively sloped response function that crosses the 45-degree line.

1.2.1 Both players choose unconditionally

If both players exhibit unconditional strategies, then $d_i = a_i$ and $d_{i'} = a_{i'}$. In this case,

$$\begin{aligned} n = 1, \quad & x_n = \tilde{x}_i \\ & y_n = a_{i'} \\ n \geq 2, \quad & x_n = a_i \\ & y_n = a_{i'}. \end{aligned} \tag{31}$$

1.2.2 First mover chooses conditionally, second mover unconditionally

This case means that $d_i \neq a_i$, $d_{i'} = a_{i'}$, and

$$\begin{aligned} n = 1, \quad & x_n = \tilde{x}_i \\ & y_n = a_{i'} \\ n \geq 2, \quad & x_n = a_i + a_{i'}(d_i - a_i) \\ & y_n = a_{i'}. \end{aligned} \tag{32}$$

1.2.3 First mover chooses unconditionally, second mover conditionally

When $d_i = a_i$ and $d_{i'} \neq a_{i'}$,

$$\begin{aligned} n = 1, \quad & x_n = \tilde{x}_i \\ & y_n = a_{i'} + \tilde{x}_i(d_{i'} - a_{i'}) \\ n \geq 2, \quad & x_n = a_i \\ & y_n = a_{i'} + a_i(d_{i'} - a_{i'}). \end{aligned} \tag{33}$$

1.2.4 Both players perfectly reciprocate

In this case, $d_i = 1$, $a_i = 0$, $d_{i'} = 1$, and $a_{i'} = 0$. The first mover's initial transfer persists indefinitely. That is, $\forall n \geq 1$, $x_n = y_n = \tilde{x}_i$.

1.2.5 Both players perfectly anti-reciprocate

When both players are perfect anti-reciprocators, $d_i = 0$, $a_i = 1$, $d_{i'} = 0$, and $a_{i'} = 1$. In this case, $\forall n \geq 1$, $x_n = \tilde{x}_i$ and $y_n = 1 - \tilde{x}_i$.

1.2.6 First mover perfectly reciprocates, second mover perfectly anti-reciprocates

For this mix of player types, $d_i = 1$, $a_i = 0$, $d_{i'} = 0$, and $a_{i'} = 1$. Player choices simplify to

$$\begin{aligned} n \in \{1, 3, \ldots\}, \quad & x_n = \tilde{x}_i \\ & y_n = 1 - \tilde{x}_i \\ n \in \{2, 4, \ldots\}, \quad & x_n = 1 - \tilde{x}_i \\ & y_n = \tilde{x}_i. \end{aligned} \tag{34}$$

1.2.7 First mover perfectly anti-reciprocates, second mover perfectly reciprocates

Under this pairing, $d_i = 0$, $a_i = 1$, $d_{i'} = 1$, and $a_{i'} = 0$. Player choices follow

$$\begin{aligned} n \in \{1, 3, \ldots\}, \quad & x_n = \tilde{x}_i \\ & y_n = \tilde{x}_i \\ n \in \{2, 4, \ldots\}, \quad & x_n = 1 - \tilde{x}_i \\ & y_n = 1 - \tilde{x}_i. \end{aligned} \tag{35}$$

1.2.8 Both players reciprocate, at least one imperfectly

This case means that $(d_i - a_i) \in (0, 1]$, $(d_{i'} - a_{i'}) \in (0, 1]$ and $(d_i - a_i)(d_{i'} - a_{i'}) \in (0, 1)$. With these constraints, one can show that, for $l \in \{1, 2\}$,

$$\sum_{k=0}^{n-l} (d_i - a_i)^k (d_{i'} - a_{i'})^k = \frac{1 - (d_i - a_i)^{n-l+1}(d_{i'} - a_{i'})^{n-l+1}}{1 - (d_i - a_i)(d_{i'} - a_{i'})}.$$

As a result,

$$\begin{aligned} \lim_{n\to\infty} x_n &= \frac{a_i + a_{i'}(d_i - a_i)}{1 - (d_i - a_i)(d_{i'} - a_{i'})} \\ \lim_{n\to\infty} y_n &= \frac{a_i(d_{i'} - a_{i'}) + a_{i'}}{1 - (d_i - a_i)(d_{i'} - a_{i'})}. \end{aligned} \tag{36}$$

One can easily verify that, if $d_i < 1$ or $d_{i'} < 1$, $\lim_{n\to\infty} x_n < 1$ and $\lim_{n\to\infty} y_n < 1$. In other words, if at least one of the two agents is an ambiguous reciprocator, both will converge to an interior level of cooperation as their relationship continues.

1.2.9 Both players anti-reciprocate, at least one imperfectly

This scenario requires that $(d_i - a_i) \in [-1, 0)$, $(d_{i'} - a_{i'}) \in [-1, 0)$, and $(d_i - a_i)(d_{i'} - a_{i'}) \in (0, 1)$. The limiting results from directly above (36), where both reciprocate, at least one imperfectly, also hold in the present case.

1.2.10 First mover reciprocates, second mover anti-reciprocates, at least one imperfectly

Here, $(d_i - a_i) \in (0, 1]$, $(d_{i'} - a_{i'}) \in [-1, 0)$, and $(d_i - a_i)(d_{i'} - a_{i'}) \in (-1, 0)$. To understand limiting behaviours, note that

$$\begin{aligned}\sum_{k=0}^{\infty}(d_i - a_i)^k(d_{i'} - a_{i'})^k &= \{1 + (d_i - a_i)(d_{i'} - a_{i'})\}\sum_{k=1}^{\infty}(d_i - a_i)^{2k-2}(d_{i'} - a_{i'})^{2k-2} \\ &= \frac{1 + (d_i - a_i)(d_{i'} - a_{i'})}{1 - (d_i - a_i)^2(d_{i'} - a_{i'})^2}.\end{aligned} \tag{37}$$

By extension, for $l \in \{1, 2\}$,

$$\sum_{k=0}^{n-l}(d_i - a_i)^k(d_{i'} - a_{i'})^k = \frac{(1 + (d_i - a_i)(d_{i'} - a_{i'}))\left\{1 - (d_i - a_i)^{n-l+1}(d_{i'} - a_{i'})^{n-l+1}\right\}}{1 - (d_i - a_i)^2(d_{i'} - a_{i'})^2}. \tag{38}$$

Player choices converge as follows,

$$\begin{aligned}\lim_{n\to\infty} x_n &= \frac{(a_i + a_{i'}(d_i - a_i))\{1 + (d_i - a_i)(d_{i'} - a_{i'})\}}{1 - (d_i - a_i)^2(d_{i'} - a_{i'})^2} \\ \lim_{n\to\infty} y_n &= \frac{(a_i(d_{i'} - a_{i'}) + a_{i'})\{1 + (d_i - a_i)(d_{i'} - a_{i'})\}}{1 - (d_i - a_i)^2(d_{i'} - a_{i'})^2}.\end{aligned} \tag{39}$$

1.2.11 First mover anti-reciprocates, second mover reciprocates, at least one imperfectly

This scenario requires that $(d_i - a_i) \in [-1, 0)$, $(d_{i'} - a_{i'}) \in (0, 1]$, and $(d_i - a_i)(d_{i'} - a_{i'}) \in (-1, 0)$. The limiting results from directly above (39), where the first mover reciprocates and the second mover anti-reciprocates, at least one imperfectly, also hold here.

1.2.12 Escalating monomorphisms on the brink of susceptibility to invasion

In the analysis that follows, we will examine populations that are monomorphic apart from a rare mutant, and we will ask if and when the resident strategy is susceptible to invasion by the mutant.

Importantly, to do so, we will not allow for any arbitrary mutation in the three-dimensional strategy space. Such an analysis is algebraically cumbersome, and it leads to few insights. Moreover, a three-dimensional strategy space is less amenable to the graphical approach we used above for two-dimensional strategies. For this reason, we restrict attention to mutations in single dimensions. This approach is broadly consistent with the uncorrelated mutations we used in our agent-based simulations (§ 2) because, if mutations are uncorrelated across dimensions, simultaneous mutations in more than one dimension are exceedingly unlikely. More importantly, as we show, analysing mutations in single dimensions is enough to identify the critical vulnerability of reciprocal strategies in strategy spaces with more than two dimensions.

Posit a monomorphic population of individuals of type R. Individuals of this type have strategies characterised by $\tilde{x}$, a, and d, where $(d-a) \in [0,1)$. The general solution for choices in a pair simplifies to

$$\begin{aligned} n=1, \quad & x_n = \tilde{x} \\ & y_n = a + (d-a)\tilde{x} \\ n \geq 2, \quad & x_n = (d-a)^{2n-2}\tilde{x} + a\sum_{k=0}^{2n-3}(d-a)^k \\ & y_n = (d-a)^{2n-1}\tilde{x} + a\sum_{k=0}^{2n-2}(d-a)^k. \end{aligned} \tag{40}$$

The expected fitness values of the common type, conditional on role in the game, are

$$\begin{aligned}
\mathbb{E}[\Pi_1(R;R)] &= \omega_0 + 1 - \tilde{x} + b\{a + (d-a)\tilde{x}\} + \sum_{n=2}^{\infty} \delta^{n-1}\{1 - x_n + by_n\} \\
&= \omega_0 + \frac{1}{1-\delta} - \tilde{x} + b\{a + (d-a)\tilde{x}\} + \frac{a\delta(b-1)}{(1-\delta)(1-(d-a))} \\
&\qquad + \frac{\tilde{x}\delta(d-a)^2\{b(d-a)-1\}}{1-\delta(d-a)^2} \\
&\qquad + \frac{a\delta(d-a)^2\{1-b(d-a)\}}{(1-(d-a))\{1-\delta(d-a)^2\}} \\
\mathbb{E}[\Pi_2(R;R)] &= \omega_0 + 1 - a - (d-a)\tilde{x} + b\tilde{x} + \sum_{n=2}^{\infty} \delta^{n-1}\{1 - y_n + bx_n\} \\
&= \omega_0 + \frac{1}{1-\delta} - a - (d-a)\tilde{x} + b\tilde{x} + \frac{a\delta(b-1)}{(1-\delta)(1-(d-a))} \\
&\qquad + \frac{\tilde{x}\delta(d-a)^2\{b-(d-a)\}}{1-\delta(d-a)^2} \\
&\qquad + \frac{a\delta(d-a)^2\{(d-a)-b\}}{(1-(d-a))\{1-\delta(d-a)^2\}}.
\end{aligned} \tag{41}$$

To simplify the notation, define $D = d - a$. The unconditional expected fitness value for individuals of the resident type is

$$\begin{aligned}
W(R) &= (1/2)\left\{\mathbb{E}[\Pi_1(R;R)] + \mathbb{E}[\Pi_2(R;R)]\right\} \\
&= \omega_0 + \frac{1}{1-\delta} + \frac{(b-1)\{a + (d-a)\tilde{x} + \tilde{x}\}}{2} + \frac{a\delta(b-1)}{(1-\delta)(1-(d-a))} \\
&\qquad + \frac{\tilde{x}\delta(d-a)^2(b-1)(1+(d-a))}{2\{1-\delta(d-a)^2\}} - \frac{a\delta(d-a)^2(b-1)(1+(d-a))}{2(1-(d-a))\{1-\delta(d-a)^2\}} \\
&= \omega_0 + \frac{1}{1-\delta} + \left(\frac{b-1}{2}\right)\left\{\frac{a + a\delta(1+D) + a\delta D}{(1-\delta)(1-\delta D^2)} + \frac{\tilde{x}(1+D)}{1-\delta D^2}\right\}.
\end{aligned} \tag{42}$$

Posit a rare mutant, R', with a strategy characterised by initial transfer $\tilde{z}$, left intercept g, and right intercept h. When this mutant meets an individual of the resident type, and the mutant is in the role

of first mover, the general solution for choices is

$$\begin{aligned}
n = 1, \quad & x_n = \tilde{z} \\
& y_n = a + (d-a)\tilde{z} \\
n \geq 2, \quad & x_n = (d-a)^{n-1}(h-g)^{n-1}\tilde{z} \\
& \qquad + a\sum_{k=0}^{n-2}(d-a)^k(h-g)^{k+1} + g\sum_{k=0}^{n-2}(d-a)^k(h-g)^k \\
& y_n = (d-a)^n(h-g)^{n-1}\tilde{z} + a\sum_{k=0}^{n-1}(d-a)^k(h-g)^k + g\sum_{k=0}^{n-2}(d-a)^{k+1}(h-g)^k.
\end{aligned} \tag{43}$$

The expected fitness value of the mutant, conditional on role as first mover, is

$$\begin{aligned}
\mathbb{E}[\Pi_1(R';R)] &= \omega_0 + 1 - \tilde{z} + b\{a + (d-a)\tilde{z}\} + \sum_{n=2}^{\infty}\delta^{n-1}\{1 - x_n + by_n\} \\
&= \omega_0 + \frac{1}{1-\delta} - \tilde{z} + b\{a + (d-a)\tilde{z}\} + \frac{\tilde{z}\delta(d-a)(h-g)\{b(d-a)-1\}}{1-\delta(d-a)(h-g)} \\
&\qquad + \frac{\delta\{g(b(d-a)-1) + a(b-(h-g))\}}{(1-\delta)(1-(d-a)(h-g))} \\
&\qquad - \frac{\delta(d-a)(h-g)(g+a(h-g))\{b(d-a)-1\}}{(1-(d-a)(h-g))\{1-\delta(d-a)(h-g)\}}.
\end{aligned} \tag{44}$$

When the mutant is in the role of second mover, the general solution for the pair is

$$\begin{aligned}
n = 1, \quad & x_n = \tilde{x} \\
& y_n = g + (h-g)\tilde{x} \\
n \geq 2, \quad & x_n = (d-a)^{n-1}(h-g)^{n-1}\tilde{x} \\
& \qquad + a\sum_{k=0}^{n-2}(d-a)^k(h-g)^k + g\sum_{k=0}^{n-2}(d-a)^{k+1}(h-g)^k \\
& y_n = (d-a)^{n-1}(h-g)^n\tilde{x} + a\sum_{k=0}^{n-2}(d-a)^k(h-g)^{k+1} + g\sum_{k=0}^{n-1}(d-a)^k(h-g)^k.
\end{aligned} \tag{45}$$

The expected fitness value of the mutant, conditional on role as second mover, is

$$\begin{aligned}\mathbb{E}[\Pi_2(R';R)] &= \omega_0 + 1 - g - (h-g)\tilde{x} + b\tilde{x} + \sum_{n=2}^{\infty}\delta^{n-1}\{1 - y_n + bx_n\} \\ &= \omega_0 + \frac{1}{1-\delta} - g - (h-g)\tilde{x} + b\tilde{x} + \frac{\tilde{x}\delta(d-a)(h-g)\{b-(h-g)\}}{1-\delta(d-a)(h-g)} \\ &\quad + \frac{\delta\{b(a+g(d-a)) - a(h-g) - g\}}{(1-\delta)(1-(d-a)(h-g))} \\ &\quad + \frac{\delta(d-a)(h-g)(a+g(d-a))\{(h-g)-b\}}{(1-(d-a)(h-g))\{1-\delta(d-a)(h-g)\}}.\end{aligned} \tag{46}$$

To simplify the notation, let $H = h - g$. The unconditional expected fitness value for the rare mutant is

$$\begin{aligned}W(R') &= (1/2)\left\{\mathbb{E}[\Pi_1(R';R)] + \mathbb{E}[\Pi_2(R';R)]\right\} \\ &= \omega_0 + \frac{1}{1-\delta} + \frac{b\{a + D\tilde{z} + \tilde{x}\} - \tilde{z} - g - H\tilde{x}}{2} \\ &\quad + \frac{\delta DH\{\tilde{z}(bD-1) + \tilde{x}(b-H)\}}{2(1-\delta DH)} + \frac{\delta\{g(bD-1) + ab - aH\}}{(1-\delta)(1-DH)} \\ &\quad + \frac{\delta DH\{2aH - 2bgD + (1+DH)(g-ab)\}}{2(1-DH)(1-\delta DH)}.\end{aligned} \tag{47}$$

Assume the rare mutation only involves the initial transfer if first mover, and so $g = a$ and $h = d$. In this case, $W(R') > W(R)$ if $(\tilde{x}-\tilde{z})(1-b(d-a)) > 0$. This means that, if $b(d-a)$ is sufficiently small, the rare mutant will invade if its initial transfer value is less than that of the resident type. If $b(d-a)$ is sufficiently large, the rare mutant will invade if its initial transfer value is more than that of the resident type.

Now assume the rare mutation involves only the response function, and $\tilde{z} = \tilde{x}$. In this case,

$$\begin{aligned}W(R') &= \omega_0 + \frac{1}{1-\delta} + \frac{\tilde{x}\{b(1+D) - (1+H)\}}{2(1-\delta DH)} \\ &\quad + \frac{(ab-g)(1+\delta)}{2(1-\delta)(1-\delta DH)} + \frac{\delta(bgD - aH)}{(1-\delta)(1-\delta DH)}.\end{aligned} \tag{48}$$

By extension,

$$\begin{aligned} W(R') - W(R) &= \frac{\tilde{x}(D-H)\{1-b\delta D(1+D)+\delta D\}}{2(1-\delta DH)(1-\delta D^2)} + \frac{g\{2b\delta D-(1+\delta)\}}{2(1-\delta)(1-\delta DH)} \\ &\quad + \frac{a\{\delta(D-H)(2-b\delta D-bD)+(1-\delta DH)(1-2b\delta D+\delta)\}}{2(1-\delta)(1-\delta DH)(1-\delta D^2)} \\ &= \frac{(D-H)\{\delta(1-bD)(\tilde{x}D(1-\delta)+a)+(1-b\delta D)(\tilde{x}(1-\delta)+a\delta)\}}{2(1-\delta)(1-\delta DH)(1-\delta D^2)} \\ &\quad + \frac{(1-2b\delta D+\delta)\{a-g-\delta D(aH-gD)\}}{2(1-\delta)(1-\delta DH)(1-\delta D^2)}. \end{aligned} \tag{49}$$

If $W(R') - W(R) > 0$, the rare mutant has strictly higher fitness and invades the population of resident escalators.

Unfortunately, extracting illuminating results from (49) is difficult. Nonetheless, we can make some progress if we consider two special cases. First, we restrict attention to an analysis of escalation, and so we assume $0 < a, g < 1$, $a \neq g$, and $d = h = 1$. Moreover, because all forms of escalation are equivalent when $\tilde{x} = 1$, we also assume $\tilde{x} < 1$. In this case, one can easily verify that the following degree of escalation strictly dominates local deviations in terms of escalation,

$$\hat{D}_1 = \frac{\delta(1-b) + \sqrt{\delta^2(1-b)^2 + 4b\delta}}{2b\delta}. \tag{50}$$

This result is exactly the same as the result with two-dimensional strategies (28), where $\hat{\xi}_1 = 1 - \hat{\alpha}_1$ in two dimensions is equivalent to $\hat{D}_1 = 1 - \hat{a}_1$ in three dimensions.

Second, we consider the special case in which a mutant ambiguous reciprocator appears in a population of escalating reciprocators. Specifically, we consider $0 < a = g < h < d = 1$. In this case, $W(R') - W(R) > 0$ if and only if $f(a) > 0$, where

$$\begin{aligned} f(a) &= \delta(1-bD)(\tilde{x}D(1-\delta)+a(1+\delta D)) + (1-b\delta D)(\tilde{x}(1-\delta)+a\delta(1+D)) \\ &= \delta(1-b(1-a))(\tilde{x}(1-a)(1-\delta)+a(1+\delta(1-a))) \\ &\quad + (1-b\delta(1-a))(\tilde{x}(1-\delta)+a\delta(1+(1-a))). \end{aligned} \tag{51}$$

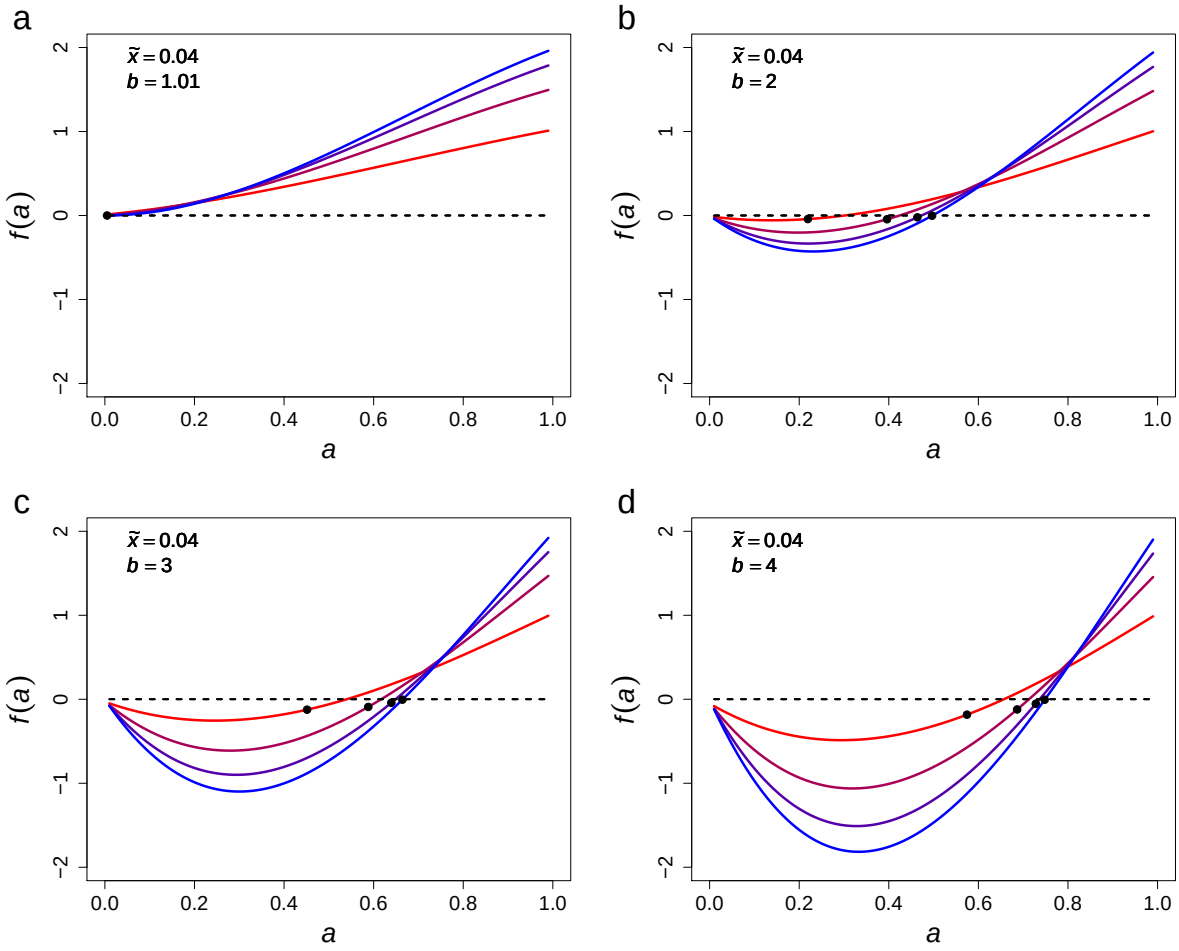
If $\delta = 0$, $f(a) = \tilde{x}$, and the mutant ambiguous reciprocator has strictly higher fitness than the resident escalators if and only if $\tilde{x} > 0$. This happens because the two strategies are equivalent as first movers, but given that $a = g$ the ambiguous reciprocator does better as second mover if the first mover cooperates to some degree. If $\delta > 0$ and if $1/b > D$, all terms in (51) are positive, and $W(R') - W(R) > 0$. This result indicates that, when interactions are repeated, a sufficiently strong degree of escalation is always susceptible to invasion by a mutant ambiguous reciprocator of the type under consideration. In contrast, if $\delta > 0$ and if $1/b < \delta D$, all terms in (51) are negative, and $W(R') - W(R) < 0$. This result shows that when interactions are repeated, a population of sufficiently weak escalators resists invasion by the ambiguous reciprocator we consider here as long as the continuation probability (δ) and gains from cooperation (b) are sufficiently large.

A more complicated case occurs when $\delta D < 1/b < D$. When these conditions are satisfied, the $(1 - bD)$ term in $f(a)$ is negative, while the $(1 - b\delta D)$ term is positive. Consequently, the sign of $f(a)$ depends in a complicated way on all parameter values and the resident value of a. Interestingly, one can easily verify that $\delta \hat{D}_1 < 1/b < \hat{D}_1$. As a result, when a population of escalators with $\tilde{x} < 1$ is at the degree of escalation given by $\hat{a}_1 = 1 - \hat{D}_1$, the population is always in this relatively complicated case. Supplementary Figures 9 - 12 show $f(a)$ under a wide variety of values for the parameters $\tilde{x}$, b, and δ. As these figures show, when $\hat{a}_1 \in (0, 1)$ exists, $f(\hat{a}_1)$ is always either (i) positive or (ii) negative but very close to zero.

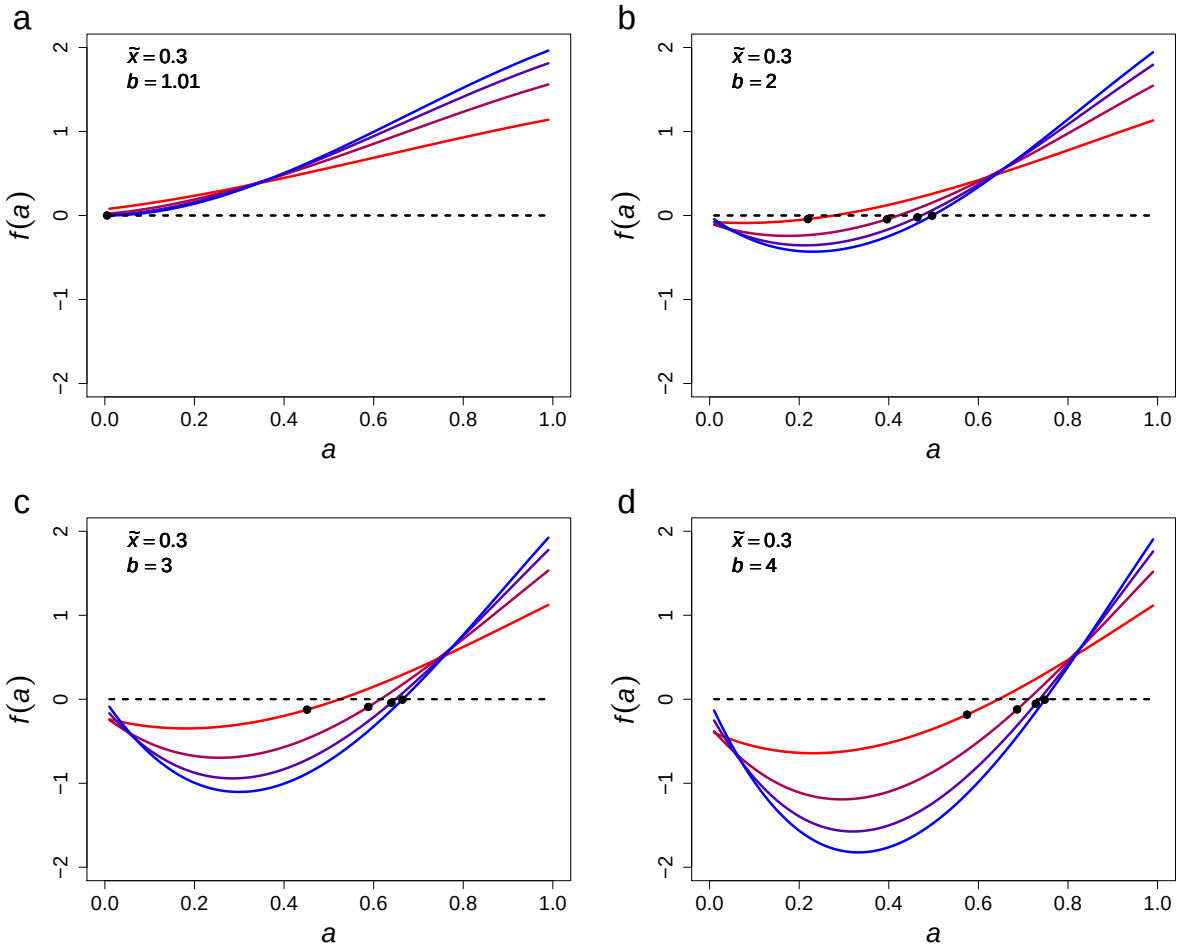
These results suggest the following. If conditions favour cooperation and ambiguously reciprocal strategies are excluded by fiat, we expect the population to evolve toward strategies in which the initial transfer is cooperative and the degree of escalation is in the neighbourhood of $\hat{a}_1$ (Supplementary Figures 6-8). However, when ambiguous reciprocity is allowed, this degree of escalation is typically either (i) susceptible to invasion by an ambiguously reciprocal strategy or (ii) resistant to invasion but extremely close to some other degree of escalation that is susceptible. In

the latter case, a finite population can easily drift toward some degree of escalation susceptible to invasion. Indeed, when initial transfers are highly cooperative, all forms of escalation are extremely similar in terms of behaviour and fitness values. Selection on the degree of escalation should be correspondingly weak and thus allow quite a bit of movement in the vicinity of $\hat{a}_1$.

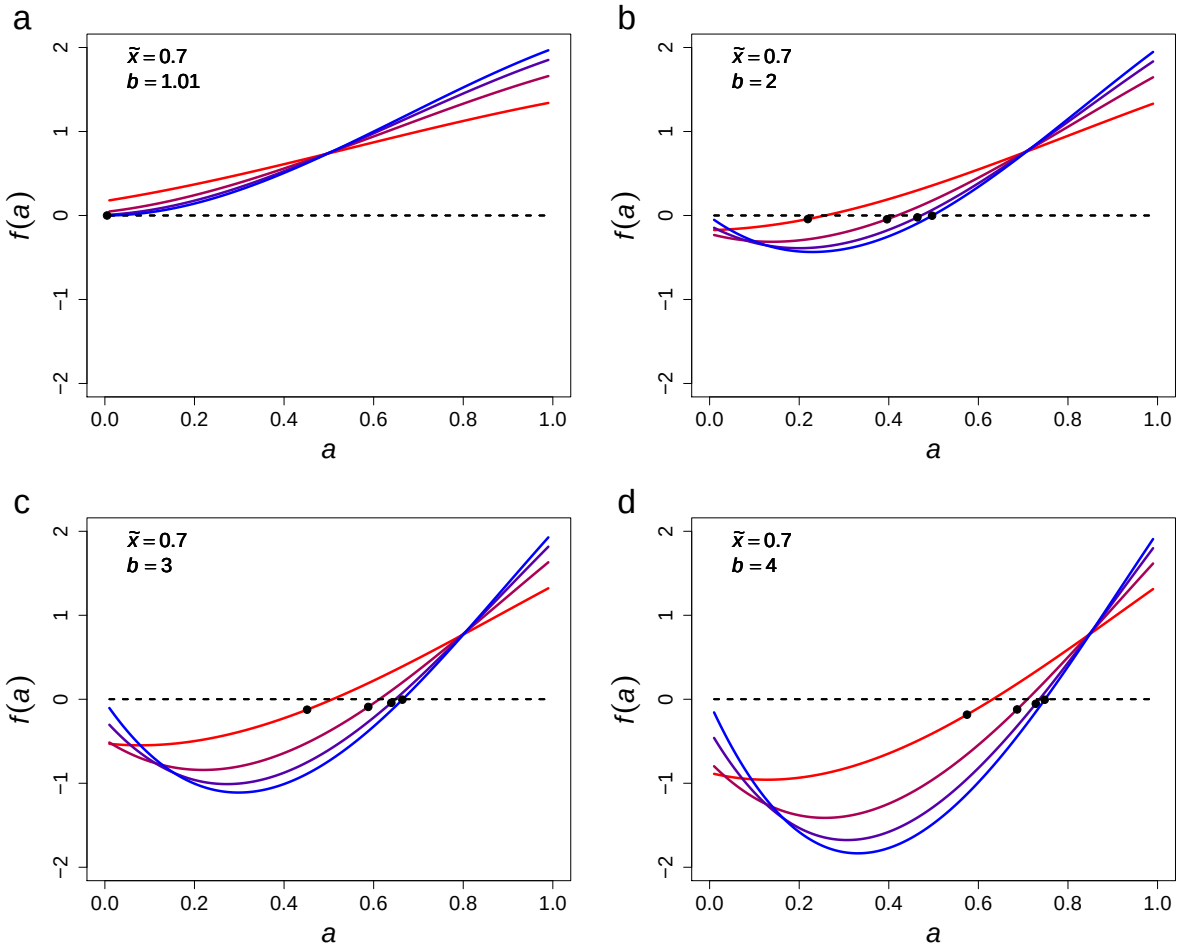
Supplementary Figures 13-14 show simulation results (see § 2) demonstrating dynamics of this sort. Specifically, under conditions quite favourable to the evolution of escalating reciprocity, escalating strategies quickly proliferate. Once this happens, however, ambiguous forms of reciprocity start to invade and destabilise escalating strategies. As ambiguous strategies become slightly more common, uncooperative de-escalating strategies spread. The end result is a population comprised largely of unconditionally selfish individuals and de-escalating reciprocators. This combination, of course, generates very little cooperation. Ambiguous strategies do not persist, but allowing them to arise via mutation has a profoundly destabilising impact on the evolution of cooperative forms of reciprocity under repeated interactions. As we show below and in the main paper (Fig. 2), this basic result continues to hold as we increase the dimensionality of strategy space to four dimensions. Importantly, four dimensions allows response functions to be sigmoidal (§ 2.3), which indicates that the result in three dimensions is not an artefact of the fact that response functions can only cross the 45-degree line from above.



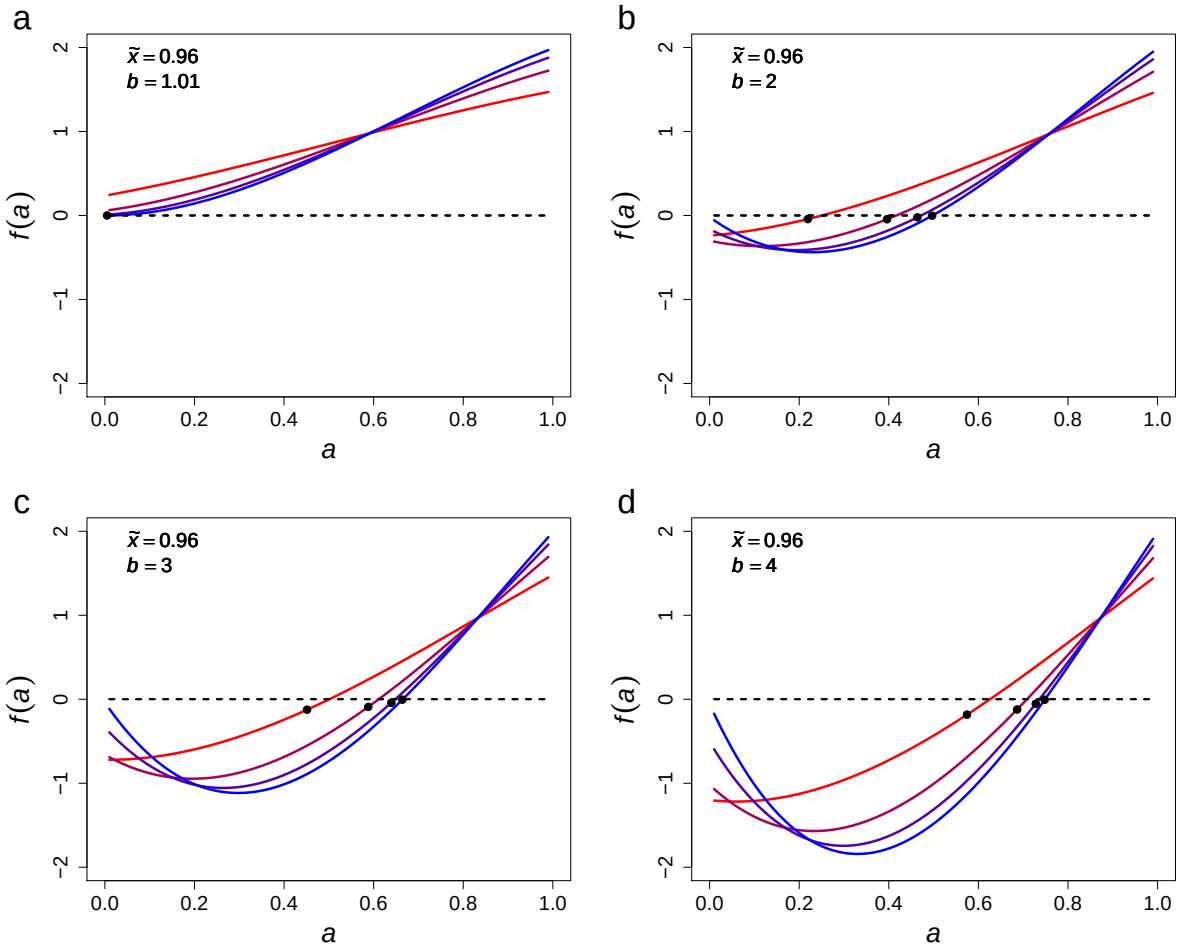
Supplementary Figure 9 | Invasion conditions for ambiguous reciprocators. The graphs show $f(a)$ when $0 < a = g < h < d = 1$. As explained in § 1.2.12, in this specific case ambiguous reciprocators have strictly higher expected fitness than resident escalators if and only if $f(a) > 0$. For all panels, $\tilde{x} = 0.04$. Panels **a** - **d** differ in terms of their values for b. Within each panel, $f(a)$ is shown for $\delta = 0.04$ (red), $\delta = 0.3$ (red-purple), $\delta = 0.7$ (purple-blue), and $\delta = 0.96$ (blue). If a combination of parameter values yields $\hat{a}_1 \in (0, 1)$, the point $(\hat{a}_1, f(\hat{a}_1))$ is shown in black. The points $(\hat{a}_1, f(\hat{a}_1))$ reveal that if a population of escalators is monomorphic at $\hat{a}_1$, a degree of escalation locally resistant to invasion by other escalating strategies, the population will often be in the immediate vicinity of a nearby degree of escalation that is vulnerable to invasion by ambiguous reciprocators. In other words, the population is on the brink of susceptibility to invasion. In a finite population with high initial transfer values, drift with respect to escalation should be relatively important, suggesting populations can readily drift from $\hat{a}_1$ to some nearby degree of invasion susceptible to ambiguously reciprocal strategies.



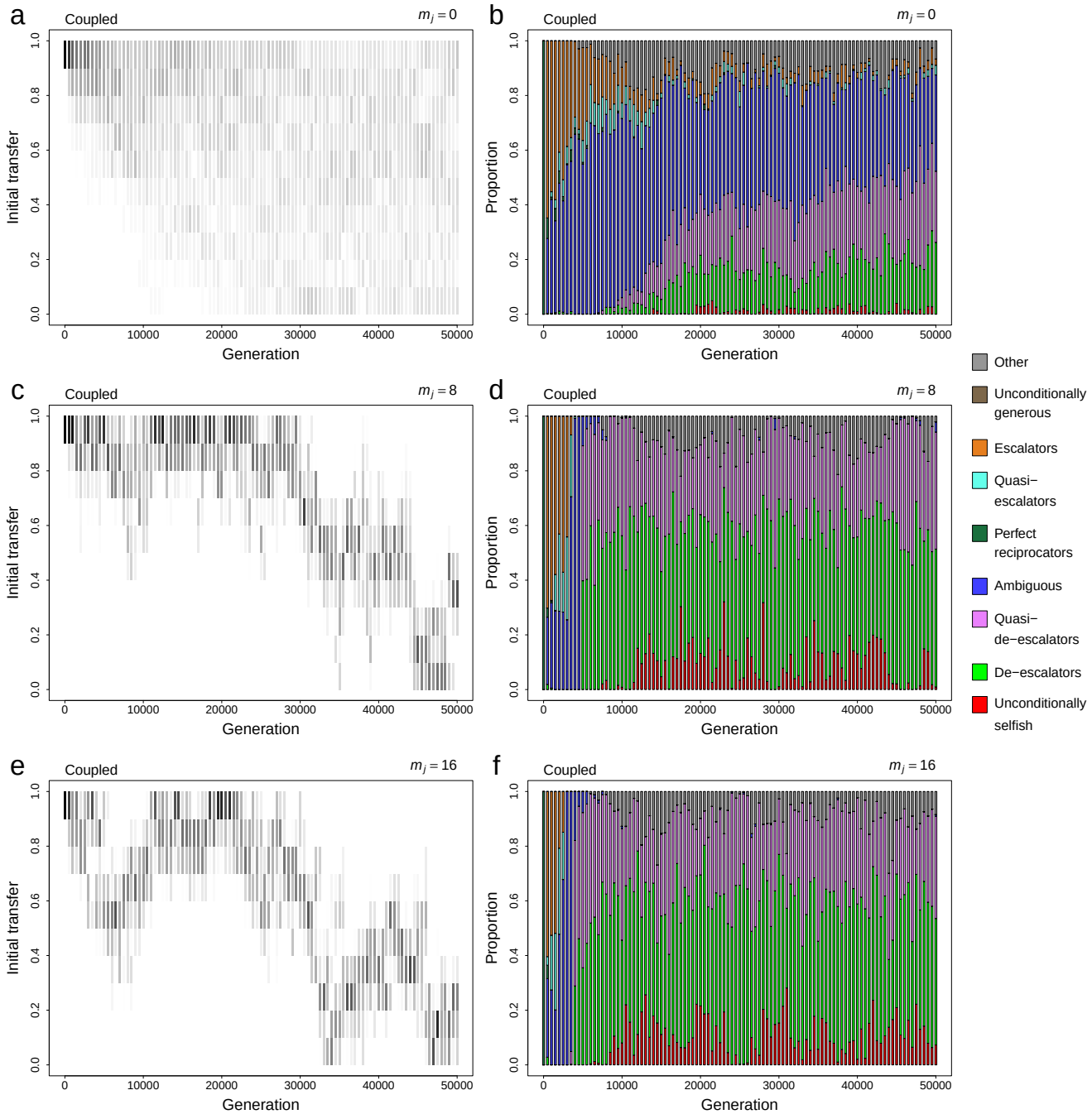
Supplementary Figure 10 | Invasion conditions for ambiguous reciprocators. The graphs show $f(a)$ when $0 < a = g < h < d = 1$. As explained in § 1.2.12, in this specific case ambiguous reciprocators have strictly higher expected fitness than resident escalators if and only if $f(a) > 0$. For all panels, $\tilde{x} = 0.3$. Panels **a** - **d** differ in terms of their values for b. Within each panel, $f(a)$ is shown for $\delta = 0.04$ (red), $\delta = 0.3$ (red-purple), $\delta = 0.7$ (purple-blue), and $\delta = 0.96$ (blue). If a combination of parameter values yields $\hat{a}_1 \in (0, 1)$, the point $(\hat{a}_1, f(\hat{a}_1))$ is shown in black. The points $(\hat{a}_1, f(\hat{a}_1))$ reveal that if a population of escalators is monomorphic at $\hat{a}_1$, a degree of escalation locally resistant to invasion by other escalating strategies, the population will often be in the immediate vicinity of a nearby degree of escalation that is vulnerable to invasion by ambiguous reciprocators. In other words, the population is on the brink of susceptibility to invasion. In a finite population with high initial transfer values, drift with respect to escalation should be relatively important, suggesting populations can readily drift from $\hat{a}_1$ to some nearby degree of invasion susceptible to ambiguously reciprocal strategies.



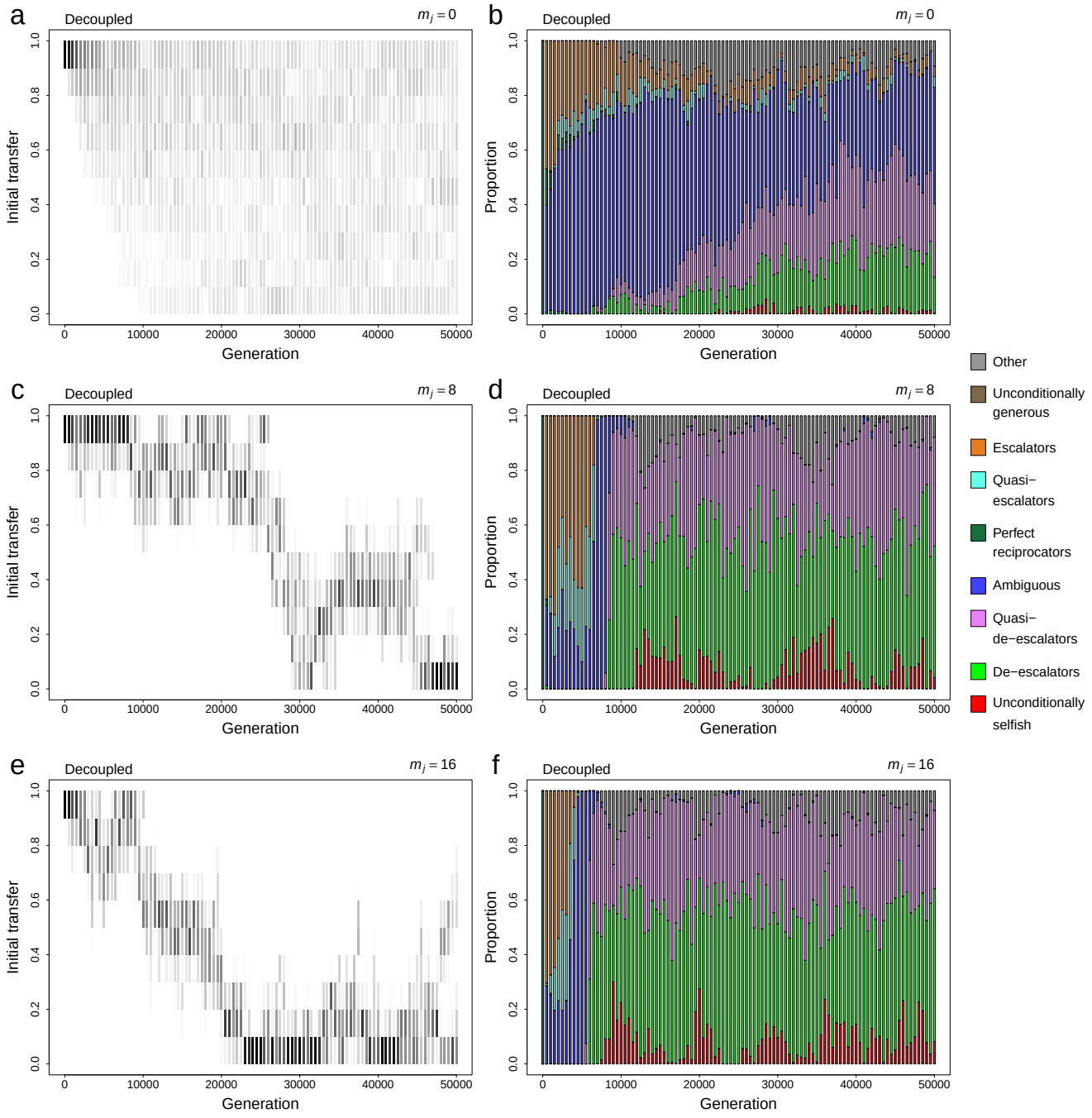
Supplementary Figure 11 | Invasion conditions for ambiguous reciprocators. The graphs show $f(a)$ when $0 < a = g < h < d = 1$. As explained in § 1.2.12, in this specific case ambiguous reciprocators have strictly higher expected fitness than resident escalators if and only if $f(a) > 0$. For all panels, $\tilde{x} = 0.7$. Panels **a** - **d** differ in terms of their values for b. Within each panel, $f(a)$ is shown for $\delta = 0.04$ (red), $\delta = 0.3$ (red-purple), $\delta = 0.7$ (purple-blue), and $\delta = 0.96$ (blue). If a combination of parameter values yields $\hat{a}_1 \in (0, 1)$, the point $(\hat{a}_1, f(\hat{a}_1))$ is shown in black. The points $(\hat{a}_1, f(\hat{a}_1))$ reveal that if a population of escalators is monomorphic at $\hat{a}_1$, a degree of escalation locally resistant to invasion by other escalating strategies, the population will often be in the immediate vicinity of a nearby degree of escalation that is vulnerable to invasion by ambiguous reciprocators. In other words, the population is on the brink of susceptibility to invasion. In a finite population with high initial transfer values, drift with respect to escalation should be relatively important, suggesting populations can readily drift from $\hat{a}_1$ to some nearby degree of invasion susceptible to ambiguously reciprocal strategies.



Supplementary Figure 12 | Invasion conditions for ambiguous reciprocators. The graphs show $f(a)$ when $0 < a = g < h < d = 1$. As explained in § 1.2.12, in this specific case ambiguous reciprocators have strictly higher expected fitness than resident escalators if and only if $f(a) > 0$. For all panels, $\tilde{x} = 0.96$. Panels **a** - **d** differ in terms of their values for b. Within each panel, $f(a)$ is shown for $\delta = 0.04$ (red), $\delta = 0.3$ (red-purple), $\delta = 0.7$ (purple-blue), and $\delta = 0.96$ (blue). If a combination of parameter values yields $\hat{a}_1 \in (0, 1)$, the point $(\hat{a}_1, f(\hat{a}_1))$ is shown in black. The points $(\hat{a}_1, f(\hat{a}_1))$ reveal that if a population of escalators is monomorphic at $\hat{a}_1$, a degree of escalation locally resistant to invasion by other escalating strategies, the population will often be in the immediate vicinity of a nearby degree of escalation that is vulnerable to invasion by ambiguous reciprocators. In other words, the population is on the brink of susceptibility to invasion. In a finite population with high initial transfer values, drift with respect to escalation should be relatively important, suggesting populations can readily drift from $\hat{a}_1$ to some nearby degree of invasion susceptible to ambiguously reciprocal strategies.



Supplementary Figure 13 | The corrosive effects of ambiguous reciprocity when the life cycle couples game play and individual selection. The graphs show the simulated (§ 2) evolution of strategies, which we have categorised into discrete types (see § 2.1.17). Panels **a**, **c**, and **e** show histograms for initial transfer values with a grey scale used to signify density. Panels **b**, **d**, and **f** show the associated distributions over types of response function. Migration rates range from zero ($m_j = 0$) to high ($m_j = 16$). The dynamics show how ambiguous reciprocity destabilises escalating strategies. This, in turn, allows de-escalating strategies to proliferate. Ambiguous strategies do not persist at high rates, but allowing them to arise has a profound effect on evolutionary dynamics.



Supplementary Figure 14 | The corrosive effects of ambiguous reciprocity when the life cycle decouples game play and individual selection. The graphs show the simulated (§ 2) evolution of strategies, which we have categorised into discrete types (see § 2.1.17). Panels **a**, **c**, and **e** show histograms for initial transfer values with a grey scale used to signify density. Panels **b**, **d**, and **f** show the associated distributions over types of response function. Migration rates range from zero ($m_j = 0$) to high ($m_j = 16$). The dynamics show how ambiguous reciprocity destabilises escalating strategies. This, in turn, allows de-escalating strategies to proliferate. Ambiguous strategies do not persist at high rates, but allowing them to arise has a profound effect on evolutionary dynamics.