



UvA-DARE (Digital Academic Repository)

Power asymmetry destabilizes reciprocal cooperation in social dilemmas

Colnaghi, Marco; Santos, Fernando P.; Van Lange, Paul A.M.; Balliet, Daniel

DOI

[10.1016/j.jtbi.2025.112106](https://doi.org/10.1016/j.jtbi.2025.112106)

Publication date

2025

Document Version

Final published version

Published in

Journal of Theoretical Biology

License

CC BY

[Link to publication](#)

Citation for published version (APA):

Colnaghi, M., Santos, F. P., Van Lange, P. A. M., & Balliet, D. (2025). Power asymmetry destabilizes reciprocal cooperation in social dilemmas. *Journal of Theoretical Biology*, 606, Article 112106. <https://doi.org/10.1016/j.jtbi.2025.112106>

General rights

It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations

If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: <https://uba.uva.nl/en/contact>, or a letter to: Library of the University of Amsterdam, Secretariat, P.O. Box 19185, 1000 GD Amsterdam, The Netherlands. You will be contacted as soon as possible.



Power asymmetry destabilizes reciprocal cooperation in social dilemmas

Marco Colnaghi^{a,*} , Fernando P. Santos^{b,1} , Paul A.M. Van Lange^{a,1}, Daniel Balliet^{a,*}

^a Department of Experimental and Applied Psychology, Institute for Brain and Behaviour Amsterdam (IBBA), Vrije Universiteit Amsterdam, Amsterdam 1081BT, the Netherlands

^b Informatics Institute, University of Amsterdam, Amsterdam, the Netherlands

ARTICLE INFO

Keywords:

Cooperation
Power
Evolutionary Game Theory
Bimatrix Games
Prisoner's Dilemma
Snowdrift Game

ABSTRACT

Direct reciprocity has been long identified as a mechanism to support the evolution of cooperation in social dilemmas. While most research on reciprocal cooperation has focused on symmetrical interactions, real world interactions often involve differences in power. Verbal theories have either claimed that power differences enhance or destabilize cooperation, indicating the need for a comprehensive theoretical model of how power asymmetries affect direct reciprocity. Here, we investigate the relationship between power and cooperation in two frequently studied social dilemmas, the prisoner's dilemma (PD) and the snowdrift game (SD). Combining evolutionary game theory and agent-based models, we demonstrate that power asymmetries are detrimental to the evolution of cooperation. Strategies that are contingent on power within an interaction provide a selective advantage in the iterated SD, but not in the iterated PD. In both games, the rate of cooperation declines as power asymmetry increases, indicating that a more egalitarian distribution of the benefits of cooperation is the prerequisite for direct reciprocity to evolve and be maintained.

1. Introduction

Social dilemmas, where individuals can increase their payoffs at the expense of collective welfare, present an important challenge to cooperation (Van Lange et al., 2015; Kollock, 1998; Dawes, 1980; Rand and Nowak, 2013). When social interactions between individuals are repeated, direct reciprocity can provide a strong incentive towards cooperative behaviour (Trivers, 1971; Axelrod, 1984; Nowak, 2006; Schmid et al., 2021; Doebeli and Hauert, 2005; Stewart and Plotkin, 2013). If the continuation probability (the likelihood of interacting again with the same individual) exceeds a critical threshold, reciprocal strategies such as Tit-For-Tat can outcompete defection, paving the way for the evolution of cooperation (Nowak, 2006; Imhof et al., 2007).

Most research on direct reciprocity in dyadic interactions has focused on symmetric games (Schmid et al., 2021; Doebeli and Hauert, 2005; Stewart and Plotkin, 2013; Imhof et al., 2007; Nowak and Sigmund, 1993; Van Veelen et al., 2012). Yet, many real-world interactions are characterized by some form of power asymmetry, where two (or more) individuals differ in how strongly they can influence each another's outcomes (Vallet et al., 2020; Scheffer et al., 2017; Piketty and Saez, 2014). Humans can readily infer and respond to power asymmetries

across social interactions (Gerpott et al., 2018; Redhead and Power, 2022; Hall et al., 2005; Fiske et al., 2007; Smith and Hofmann, 2016), and power differences are widespread in the animal kingdom as well (Kaufmann, 1983; De Vries et al., 2006; Tibbetts et al., 2022). Additionally, power asymmetry is a salient feature of interactions between humans and AI agents, a rapidly growing and increasingly relevant field of research (Zimmaro et al., 2024; Han et al., 2021; He et al., 2024; Akata et al., 2020). While the role of power asymmetry in animal contests has been investigated by several theoretical studies (Hammerstein, 1981; Gaunersdorfer et al., 1991; Maynard-Smith and Parker, 1976), its impact on cooperation is less well understood, and it is only in recent years that researchers have begun to study how power asymmetries affect direct reciprocity (Dawkins, 2010; Hauser et al., 2019; Ladret and Lessard, 2008). Here, we offer a theoretical framework for understanding why, and under what conditions, power may or may not undermine the necessary conditions for the evolution of cooperation. While our models are developed with a focus on human cooperation, the results can be readily generalized to other species.

Power can be broadly defined as asymmetric control over another person's outcomes (Fiske et al., 2007; Thibaut, 1959; Emerson, 1962). Experimental and theoretical studies on power asymmetry typically

* Corresponding authors.

E-mail addresses: m.colnaghi@vu.nl (M. Colnaghi), d.p.balliet@vu.nl (D. Balliet).

¹ These authors contributed equally to this work.

focus on factors that can lead to asymmetric control, such as differences in payoffs (Ahn et al., 2007; Beckenkamp et al., 2006; Sheposh and Gallo, 1973; Talley, 1974), endowments (Hauser et al., 2019), effectiveness of punishment (Bone et al., 2016; Nikiforakis et al., 2010), or the ability to choose one’s social partner (Hilbe et al., 2016). Empirical studies that consider differences in payoffs indicate that asymmetry can pose a barrier to the emergence of cooperation in social dilemmas (Hilbe et al., 2016; Ahn et al., 2007; Beckenkamp et al., 2006; Sheposh and Gallo, 1973). Previous theoretical research also suggests that differences in power might destabilize cooperation in social dilemmas (Dawkins, 2010) and indicates that inequality can undermine cooperation in public goods games (Hauser et al., 2019). At the same time, the effects of power differences are not unequivocal or universal; power asymmetries do not always undermine (or promote) cooperation (Bone et al., 2016; Molho et al., 2019) and power can yield different effects on cooperation in different countries (Kopelman, 2009). In fact, power hierarchies have even been considered functionally adaptive by promoting cooperation within a group (Halevy et al., 2011; Antonioni et al., 2018).

Here, we operationalise power asymmetry as the ability to provide higher or lower benefits to a partner in a social interaction (Fiske et al., 2007; French et al., 1959; Keltner et al., 2003), modelled as an iterated 2-person game. We aim to clarify how power affects the evolutionary stability of reciprocal cooperation, and under what conditions can the ability to infer power differences provide a selective advantage. We start by studying reciprocal cooperation in the iterated, asymmetric version of the simultaneous donation game, a specific form of the Prisoner’s Dilemma (PD), in a population where power differences are fixed and the population is divided in high- and low-power individuals. We derive a simple formula, which expresses the minimum continuation probability necessary to stabilize cooperation as a function of power asymmetry. We then consider a situation where power is variable (i.e., dependent on contingent factors), and introduce strategies that are conditional on power differences within interactions. Thus, we evaluate whether the ability to adapt one’s behaviour in response to power asymmetry is evolutionarily stable and provides a selective advantage in the iterated PD.

We then shift our attention to the asymmetric version of the iterated Snowdrift game (SD; also referred to as Chicken or Hawk-Dove game), another frequently studied social dilemma (Doebeli and Hauert, 2005; Maynard-Smith, 1978). In the SD, individuals can increase their payoffs by doing the opposite of what their social partner does; in line with previous studies, we refer to this behaviour as “anti-coordination” (see, for example, Bramoullé, 2007). Extending our analysis to the SD allows us to study the evolution of cooperation in a situation where differences in power can help solve collective action problems (King et al., 2009; Van Vugt et al., 2008; Pietraszewski, 2020). While the dynamics of the iterated PD become similar to a Stag Hunt (SH) (Skyrms, 2004) if the continuation probability is high enough, the dynamics of the iterated SD between reciprocal cooperators resemble a maximizing difference/harmony game (see Methods). By focusing on the iterated SD, we thus extend our analysis to all four “archetypal” games that people most frequently use to describe social interactions (Halevy et al., 2012) and are most frequently studied in the literature on social dilemmas (Santos et al., 2006; Peña and Nöldeke, 2023; Colnaghi et al., 2023).

2. Methods

2.1. Asymmetric prisoner’s dilemma

Consider a population of individuals who differ in their level of power, interacting with each other through the iterated, asymmetric simultaneous donation game. In the symmetric version of this game, each individual can incur a cost c to provide a benefit b to their social partner (Sigmund, 2010). We consider an asymmetric version of this social dilemma, where individuals can contribute a greater or smaller benefit depending on their power level, leading to a bimatrix game

(Ohtsuki, 2010) defined by the following payoff matrix:

$$\begin{array}{cc}
 & \text{Player 2} \\
 & \begin{array}{cc} C & D \end{array} \\
 \text{Player 1} & \begin{array}{cc} C & \begin{pmatrix} b_2 - c, b_1 - c & -c, b_1 \end{pmatrix} \\ D & \begin{pmatrix} b_1, -c & 0, 0 \end{pmatrix} \end{array} \end{array} \tag{1}$$

We assume that high-power individuals confer a higher benefit, $b_{HP} = (1 + \alpha)b$, and low-power individuals a lower one, $b_{LP} = (1 - \alpha)b$, with $0 \leq \alpha < 1$. The greater the asymmetry (α), the more a low-power individual’s outcome is influenced by their high-power partner’s choice to cooperate or not. By contrast, the outcome of a high-power individual is less strongly affected by their low-power partner’s decision. For simplicity, throughout the paper we refer to α as power asymmetry. This definition of power is consistent with theoretical work that suggests that the ability to allocate greater rewards is an expression of power (Fiske et al., 2007; French et al., 1959; Keltner et al., 2003). By introducing α and assuming this distribution of payoffs, we guarantee that power asymmetries do not affect the overall benefits of cooperation (i.e., the total amount of resource shared between two cooperators is always $2b$). We assume the cost of cooperation to be the same for high- and low-power individuals.

In order to study the impact of direct reciprocity, we assume that there is a continuation probability w of repeated interaction with the same individual, and individuals can either choose reciprocal cooperation (TFT) or always defect (ALLD). We limit our analysis to these two strategies as we are interested in the necessary conditions that promote the evolution of reciprocal cooperation in a population of defectors; if TFT cannot displace ALLD, no other strategy can (Nowak, 2006). Moreover, when cooperation costs are high, TFT remains a key strategy to sustain cooperation even when stochastic and longer-memory strategies are considered (Nowak et al., 2004).

In a population where power is a fixed trait, we can focus on the interactions between high and low-power individuals, and investigate whether it advantageous for two individuals to cooperate in such asymmetric interactions. Let x and y be the frequency of high- and low-power individuals, respectively, who adopt a strategy of reciprocal cooperation in asymmetric interactions. The expected payoffs for cooperators and defectors in the HP and LP groups are given by:

$$\pi_C^{HP} = \frac{(1 - \alpha)b - c}{1 - w}y - c(1 - y)$$

$$\pi_D^{HP} = (1 - \alpha)by$$

$$\pi_C^{LP} = \frac{(1 + \alpha)b - c}{1 - w}x - c(1 - x)$$

$$\pi_D^{LP} = (1 + \alpha)bx$$

Changes in x and y can be described using two coupled replicator equations:

$$\dot{x} = x(1 - x)(\pi_C^{HP} - \pi_D^{HP}) = x(1 - x) \left\{ \frac{w}{1 - w} [(1 - \alpha)b - c]y - c \right\} \tag{2.a}$$

$$\dot{y} = y(1 - y)(\pi_C^{LP} - \pi_D^{LP}) = y(1 - y) \left\{ \frac{w}{1 - w} [(1 + \alpha)b - c]x - c \right\} \tag{2.b}$$

Studying the stability of the fixed points of the dynamics (see Results section below), we evaluate the conditions that permit the evolution of reciprocal cooperation.

2.2. Asymmetric snowdrift game

The simultaneous donation game described above can be modified to assume the form of a Snowdrift (SD) game, a slightly more benign social

dilemma where cooperators and defectors can stably coexist in a population (Doebeli and Hauert, 2005; Maynard-Smith, 1978). This game represents a situation where two individuals must work together to achieve a certain outcome, such as building a shelter or freeing a road from a snowdrift, which benefits them both. The outcome is achieved if at least one individual cooperates, at a cost. If both individuals contribute, the costs are shared. Each individual would be better off if their partner would bear the whole cost of this enterprise, but incurring the cost of cooperation alone is preferable to the situation where neither player contributes to the common good.

Let $2b$ be the total payoff obtained through mutual cooperation. Suppose that the total cost of the endeavor, $2c$, can be either split between the two individuals or borne by one individual alone. As in the previous case, the reward can be either divided symmetrically when two individuals of equal power meet, or asymmetrically, if there is a power difference between the two players. This results in the bimatrix game (Ohtsuki, 2010):

$$\begin{array}{cc}
 & \text{Player 2} \\
 & \begin{array}{cc} C & D \end{array} \\
 \text{Player 1} & \begin{array}{cc} C & \left(\begin{array}{cc} b_2 - c, b_1 - c & b_2 - 2c, b_1 \end{array} \right) \\ D & \left(\begin{array}{cc} b_2, b_1 - 2c & 0, 0 \end{array} \right) \end{array}
 \end{array} \tag{3}$$

In the symmetric version of this game, both players receive the same reward b . In the asymmetric version, high-power players share with their social partners a greater proportion of the reward, $b_{HP} = (1 + \alpha)b$, and receive a smaller proportion, $b_{LP} = (1 - \alpha)b$. This specific form of SD game allows us to compare the results directly with the asymmetric PD, as the resulting payoff matrix has the same reward for mutual cooperation as payoff matrix (1). Again, we assume a continuation probability w of repeated interaction with the same partner, and consider two different strategies: reciprocal cooperation (TFT) or always defect (AllD). In populations where power is fixed, changes in frequency of cooperation in asymmetric interactions are described by the same replicator equations as in the previous section.

2.3. Evolutionary stability of strategies conditional on power

Next, we consider the more complex case of a well-mixed population where individuals play the asymmetric, iterated Prisoner’s Dilemma (PD) or Snowdrift game (SD), and have an equal probability of finding themselves in a position of high or low power. This models a situation where power depends on contingent factors (such as higher or lower resource availability, state of health, or hunting/foraging success). This scenario reflects the empirical observation that, in humans, most variation in perceived power is due to changes in situations, rather than fixed individual traits (Smith and Hofmann, 2016).

In this model, individuals can find themselves in three types of possible interactions: symmetric, low-power (when a low-power individual interacts with a high-power one), and high-power (when a high-power individual interacts with a low-power one). Assuming a well-mixed population, every individual engages in symmetric interactions with a frequency of 0.5, in interactions with a lower-power individual with a frequency of 0.25, and in interactions with a higher-power individual with a frequency of 0.25. The total payoff is averaged over a large number of such interactions. In each interaction, depending on the level of power asymmetry, an individual can choose to adopt reciprocal cooperation (TFT) or always defect (AllD).

To study whether the ability to infer differences in power can provide a selective advantage in the PD, we consider eight possible strategies: two “power-independent” strategies, where individuals always play TFT or AllD in every situation regardless of power asymmetries, and six strategies conditional on power, where individuals choose whether to play TFT or AllD depending on the specific type of interaction (symmetric or with a higher/lower-power individual). Each of the eight

strategies can then be identified as a triplet (X, Y, Z) where an individual plays X in symmetric interactions, Y when in low power and interacting with a higher-power individual, and Z when in high power and interacting with lower-power individual. From simplicity, we use the shorthand “C” to refer to reciprocal cooperation (TFT), and “D” as a shorthand for AllD. For example, the strategy (C, C, D) plays TFT in symmetric interactions and when interacting with higher-power individuals, and AllD when interacting with individuals with a lower power status. We do not consider strategies that distinguish between low-power and high-power symmetric interactions as we are mainly interested in studying the stability of strategies that depend on power asymmetries. Moreover, we focus on strategies based on AllD and TFT, as considering more than two base strategies (e.g., adding a strategy of unconditional cooperation, AllC) would significantly increase the number of power-dependent strategies to consider, and such strategies fail to invade TFT in the absence of complexity costs (see, for example, Nowak et al., 2004).

Assuming that individuals have the same probability of being in a low or high-power state, and this state can eventually change between each interaction, the total payoff of an individual adopting strategy i is $P_i = \frac{1}{2}P_i^S + \frac{1}{4}P_i^{Hf} + \frac{1}{4}P_i^L$, where P_i^S , P_i^{Hf} , and P_i^L are, respectively, the average payoff of strategy i in symmetric, high-power, and low-power interactions, which will in turn depend on the frequencies of other strategies in the population. Let P_{ji} be the average payoff of a rare mutant playing strategy j in a population where strategy i is fixed. Strategy i is evolutionarily stable if, and only if, $P_{ii} > P_{ji}$ for every other strategy j , or if $P_{ii} = P_{ji}$ and $P_{ij} > P_{ij}$ (Maynard-Smith, 1982). We evaluate numerically which strategies are evolutionarily stable for given values of α and w .

2.4. Small-mutation approximation

We also analyse the evolutionary dynamics of the strategies defined above in a finite population using a small mutation approximation, i.e., assuming that the time between the emergence of a new mutant playing a different strategy is much greater than the time it takes for a mutant to go extinct or spread to fixation (Fudenberg and Imhof, 2006). We model the evolutionary dynamics of the population as a Moran process, where at each time step an individual is randomly selected proportionally to their fitness to replace another, randomly selected individual (Moran, 1962; Traulsen et al., 2008). In line with previous theoretical work (Traulsen et al., 2008), we define the fitness of an individual adopting strategy i as $f_n(i) = e^{\beta P_n^i}$, where β denotes the strength of selection and P_n^i denotes the payoff of an individual playing strategy i in a population where there are exactly n individuals adopting strategy i (Traulsen et al., 2008). In the context of a small-mutation approximation, the population is described as dimorphic, consisting of individuals adopting strategies i and j ; therefore, the state of the population is defined by n (this contrasts with the “Exact Stationary Distribution” case below, where the population is trimorphic).

Let $x(i)$ be the frequency of strategy i . The probability that an individual adopting strategy i is selected for reproduction is $x(i)f_n(i) / \sum_k x(k)f_n(k)$, where the denominator is the average fitness of the population. The offspring of the individual selected for reproduction replaces another individual, selected at random independently of fitness; that is, the probability that an individual adopting strategy j is replaced is $x(j)/N$. Although previous works consider more sophisticated strategy update mechanisms in asymmetric iterated games (e.g., introspection dynamics; Couto et al., 2022), here we assume that strategy reproduction does not rely on complex cognitive abilities, allowing our results to fit both biological and social settings.

We model the evolution of this population as a discrete time Markov chain with transition matrix T , whose elements T_{ij} indicate the transition probability from state j (where strategy j is fixed in the population) to state i (where strategy i is fixed). The transition probability from j to i is given by the product between population size (N), mutation rate (U),

and the rate of evolution ρ_{ij} , i.e., the probability of fixation of a single mutant adopting strategy i in a population of j (Nowak et al., 2004). As the stationary distribution of a Markov chain does not change if all non-diagonal entries of the transition matrix are multiplied by the same factor, we can omit the terms N and U and write the transition coefficients as:

$$T_{ij} = \left(1 + \sum_{k=1}^{N-1} \prod_{n=1}^k \frac{f_n(j)}{f_n(i)} \right)^{-1} \quad (4)$$

As mutations from monomorphic states are considered and thus there are no absorbing states, this matrix represents an ergodic Markov chain with a unique stationary distribution, which can be calculated as the left eigenvector associated with the eigenvalue $\lambda = 1$ of the transition matrix.

2.5. Exact stationary distribution

In order to evaluate whether the small-mutation approximation is a reasonable assumption to describe the evolution of such population, we also consider the simplified case of only three strategies, for which the stationary distribution can be calculated exactly. For the PD, we considered the two power-independent strategies and one of the evolutionarily stable strategies conditional on power, (C, D, D) . For the SD, we considered the two power-independent strategies and the only evolutionarily stable strategy conditional on power, (C, C, D) . These strategies were selected as they can be evolutionarily stable for certain values of w and α in an infinite population (Figs. 2 and 4) and outcompete other conditional strategies in a finite population (Fig. 3). The population evolves according to a Moran process, as described in the previous sections, with the only difference that, every time an individual reproduces, it mutates to a different strategy with probability U . At each time step, an individual adopting strategy i is selected with probability $x(i)f_n(i)/\sum_k x(k)f(k)$. With probability $(1-U)$, their offspring adopts strategy i . With probability U , the offspring adopts a different strategy (the offspring has an equal probability of adopting any other strategy).

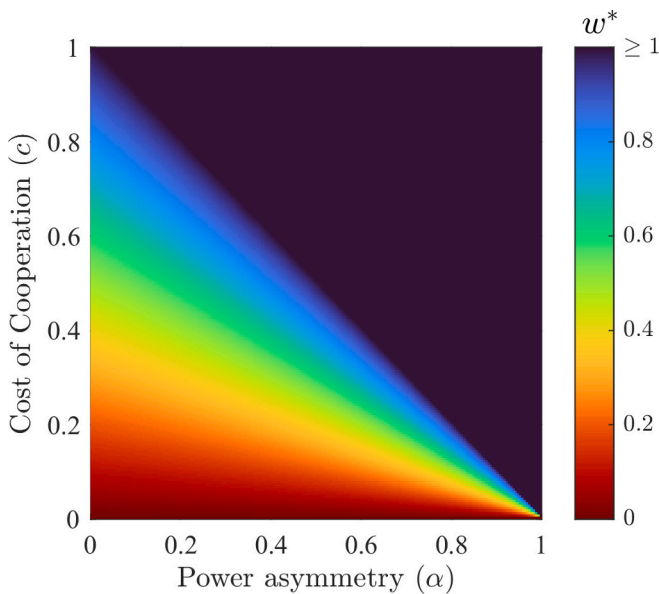


Fig. 1. Threshold continuation probability. Minimum continuation probability w^* that makes reciprocal cooperation (TFT) evolutionarily stable. Increasing power asymmetry (α) or the cost of cooperation (c) destabilizes cooperation. The dark blue color indicates the area of the parameter space where cooperation is never stable, regardless of the continuation probability. Other parameters: $b = 1$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

As previously, the offspring replaces a randomly chosen individual, independently of their fitness.

Let (k, p, q) denote a state with k individuals adopting (C, C, C) , p individuals adopting (D, D, D) and q individuals adopting (C, C, D) , and let $x = k/N$, $y = p/N$, and $z = q/N$ be the frequencies of these strategies in the population. The transition probability from (k, p, q) to $(k+1, p-1, q)$ is:

$$T_{k,p,q \rightarrow k+1,p-1,q} = xy \left(1 - \frac{3}{2}U \right) f^{CCC}(k, p, q) + \frac{U}{2}y \quad (5)$$

Where $f^{CCC}(k, p, q)$ designs the relative fitness of (C, C, C) in state (k, p, q) . Analogous formulae describe the transitions from (k, p, q) to five other possible adjacent states. The transition probabilities are calculated numerically, and the transition matrix is then used to calculate the stationary distribution. As shown in Figs. S3 and S4, up to a mutation rate of $U = 0.01$, the system spends a negligible time in the intermediate states, the pure states (where one strategy has reached fixation in the population) being the ones with the highest frequency (Figs. S3 and S4). This supports our choice of adopting the small-mutation approximation described in the previous paragraph.

3. Results

3.1. Asymmetric prisoner's dilemma

In the simple scenario where power is a fixed trait, the evolutionary dynamics are described by the replicator equations (2.a) and (2.b). It can be easily verified that the fully cooperative state $(x, y) = (1, 1)$ is a fixed point of the dynamics. This equilibrium is stable whenever the derivatives of x and y are positive in the immediate vicinity of $(1, 1)$, that is, when $\pi_C^{LP} - \pi_D^{LP} > 0$ and $\pi_C^{HP} - \pi_D^{HP} > 0$. These conditions are satisfied whenever the continuation probability exceeds the threshold:

$$w > \frac{c}{(1-\alpha)b} \quad (6)$$

This formula can be intuitively understood by bearing in mind that $(1-\alpha)b$ is the payoff of a high-power individual in asymmetric interactions (as they can allocate greater rewards, and receive smaller ones). The cost-to-benefit ratio of cooperation increases with power asymmetry, causing cooperation to collapse when it becomes too large compared to the continuation probability. When the continuation probability is smaller than this cost-to-benefit ratio, the benefit of repeated cooperation is less advantageous than exploiting one's social partner.

For $\alpha = 0$, we retrieve the well-known condition for the evolution of cooperation in the symmetrical PD, $w > c/b$ (Nowak, 2006). When $\alpha > 0$, any increase in power asymmetry leads to a higher threshold continuation probability, making it harder for cooperation to evolve (Fig. 1). As α approaches the value of 1, the minimum continuation probability necessary for cooperation to be stable diverges to infinity. This simple formula illustrates the negative impact that power asymmetry has on cooperation: even if w is high enough to sustain cooperation in the symmetrical case, increasing power inequality will eventually lead to the breakdown of cooperation. The dynamics admits another stable fixed point, $(0, 0)$ corresponding to complete lack of cooperation in asymmetric interactions. It can be easily shown that full defection is always stable, for any value of α and w . The other three fixed points of the dynamics, $(1, 0)$, $(0, 1)$, and the internal fixed point $(c(1-w)[w(b+ab-c)]^{-1}, c(1-w)[w(b-ab-c)]^{-1})$, are always unstable.

In human populations, power is generally dependent on varying situational factors (Fiske et al., 2007; Smith and Hofmann, 2016). Therefore, we next consider the more complex case of a population where power is variable, and individuals have an equal probability of finding themselves in a position of high or low power, and can adopt

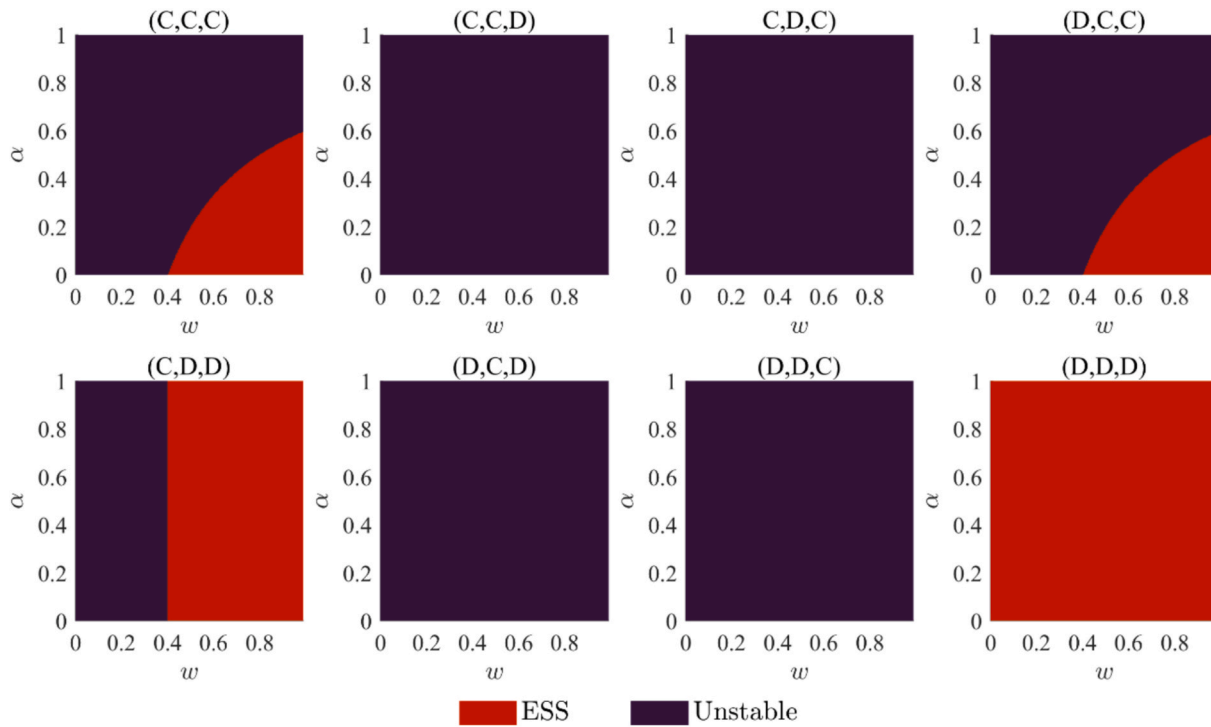


Fig. 2. Evolutionarily stable strategies in the iterated PD. Evolutionarily stable strategies in the iterated Prisoner's Dilemma for varying levels of power asymmetry (α) and continuation probability (w). Each triplet (X,Y,Z) designs a strategy where an individual plays X in symmetric interactions, Y when in low power and interacting with a higher-power individual, and Z when in high power and interacting with lower-power individual. The area of the parameter space where each strategy is an ESS is indicated in red. Other parameters: $b = 1$, $c = 0.4$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

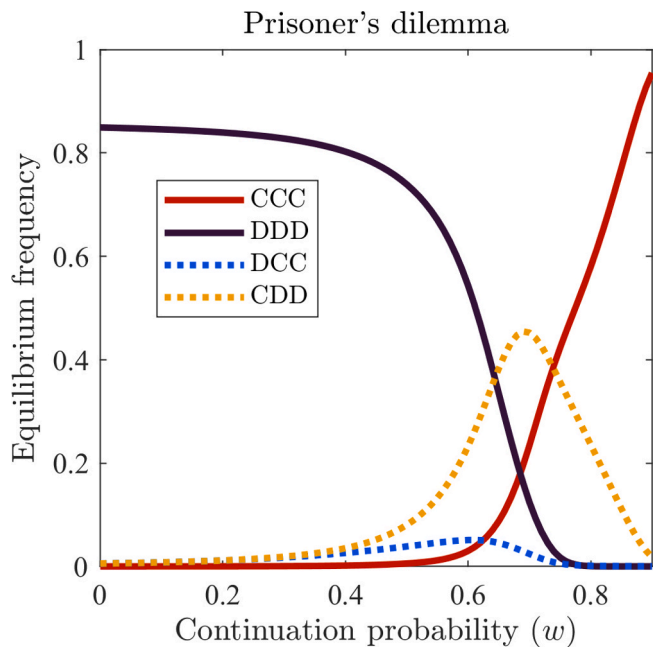


Fig. 3. Equilibrium frequency of strategies in the iterated PD. Equilibrium frequency of various strategies in a finite population playing the iterated PD, under a small-mutation approximation, as a function of the continuation probability (w). Continuous lines indicate power-independent strategies; dotted lines indicate strategies that are conditional on power differences. Of the 8 strategies studied, only the four ESSs are shown. Other parameters: $\alpha = 0.5$, $b = 1$, $c = 0.4$, $N = 500$, $\beta = 0.05$.

strategies that are contingent on the power difference within an

interaction. Using a standard evolutionary game-theoretical approach (Maynard-Smith, 1982), we evaluate under what conditions each of the strategies defined in the Methods section is evolutionarily stable (Fig. 2). In addition, we use a small-mutation approximation (Fudenberg and Imhof, 2006) to evaluate whether each of the strategies that are evolutionarily stable in an infinite population would evolve in a finite population subject to stochastic perturbations (Fig. 3).

The ESS analysis shows that, for a given level of power asymmetry, a cooperative power-independent strategy (C, C, C) is evolutionarily stable, provided that the continuation probability is high enough (Fig. 2). As power asymmetry increases, the minimum threshold that makes reciprocal cooperation advantageous increases as well, making cooperation less stable. As a single cooperative mutant can never invade an infinite population of defectors, (D, D, D) is an ESS in the whole parameter space (Fig. 2).

In addition to these two power-independent strategies, two mixed strategies, (D, C, C) and (C, D, D), are evolutionarily stable in two regions of the parameter space. However, a finite-population size analysis reveals that these two strategies almost never outperform power-independent ones (Fig. 3). (C, D, D) outperforms other strategies only for a limited range of continuation probability w and in large populations ($N = 500, 1000$; Fig. S1). (D, C, C) is never advantageous in a finite population. In principle, these strategies are stable once they spread to fixation in an infinite population; in practice, however, natural selection does not promote their spread (if not in a very narrow range of ecological conditions), making the fixation of power-independent strategies more likely. When repeated interactions are infrequent, natural selection favours defection in all situations (D, D, D) (Fig. 3). As the continuation probability increases, so does the equilibrium frequency of (C, C, C), eventually outcompeting (D, D, D): once cooperation becomes advantageous, it is favorable to cooperate in all situations, regardless of whether the interaction is symmetric or asymmetric (Fig. 3). To conclude, the ability to infer power and modify one's behaviour

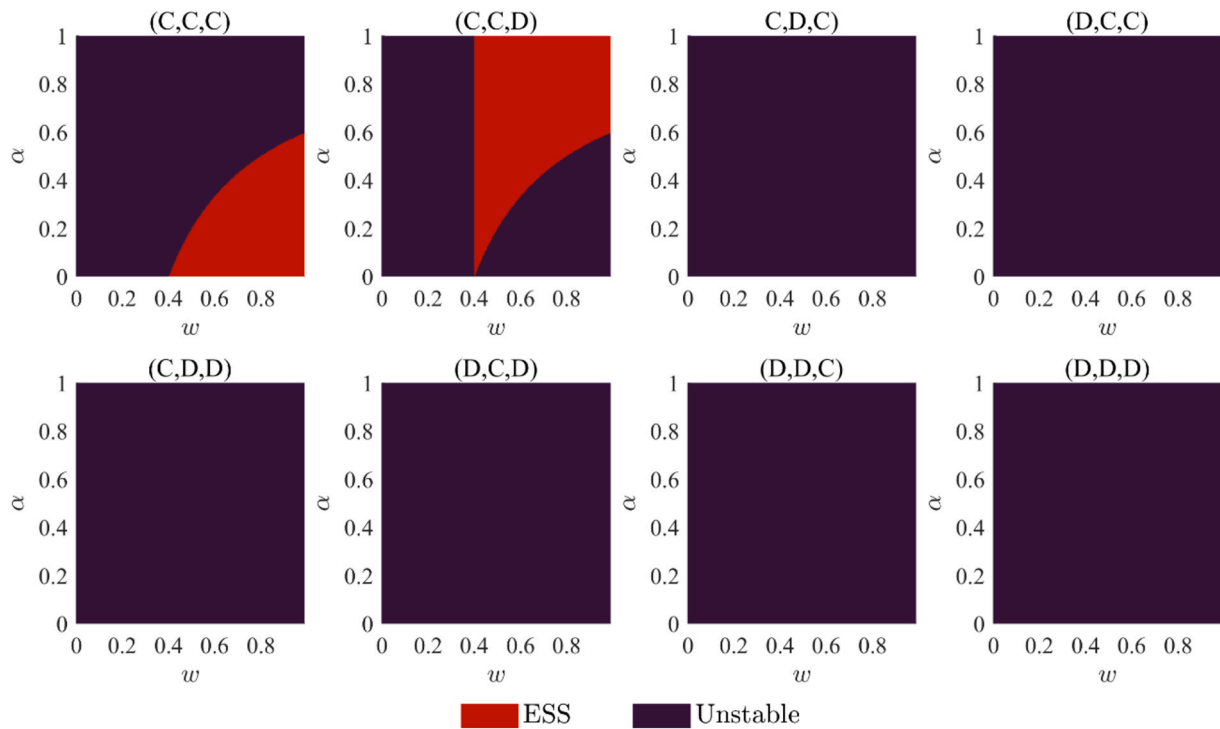


Fig. 4. Evolutionarily stable strategies in the iterated SD. Evolutionarily stable strategies in the iterated Snowdrift for varying levels of power asymmetry (α) and continuation probability (w). The area of the parameter space where each strategy is an ESS is indicated in red. Other parameters: $b = 1, c = 0.4$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

accordingly is unlikely to provide a selective advantage in the iterated asymmetric Prisoner’s Dilemma.

3.2. Asymmetric snowdrift game

In the simple case of individuals with fixed power, the stability condition for the fully cooperative equilibrium is the same as equation (6). In the more complex case of a population where individuals can find themselves in a high- or low-power position with equal probability, we consider the two power-independent and the six strategies conditional on power introduced in the Methods. Again, we evaluate which strategies are ESSs, and apply a small-mutation approximation to analyse which of the ESSs are favoured by natural selection in a finite population.

As in the case of the PD, a standard evolutionary game-theoretical analysis (Maynard-Smith, 1982) reveals that a fixed strategy of reciprocal cooperation (C, C, C) becomes evolutionarily stable once the continuation probability is high enough and power asymmetry is not too extreme (Fig. 4). As power asymmetry increases, (C, C, C) becomes unstable and is replaced by a strategy conditional on power differences, (C, C, D), which plays TFT in symmetric interactions and defects when interacting with lower-power individuals (Fig. 4). Performing a small-mutation, finite population size analysis, we find that this strategy outcompetes both power-independent strategies, (C, C, C) and (D, D, D), when the continuation probability is low (Fig. 5). As w increases, (C, C, C) becomes more favorable, and it eventually outcompetes (C, C, D) (Fig. 5). These conclusions are not affected by population size (Fig. S2). Thus, if the continuation probability does not exceed this threshold, the ability to infer power and modify one’s behavior accordingly provides a selective advantage in the iterated asymmetric SD.

3.3. Changes in cooperation frequency

Finally, we consider how the introduction of strategies conditional

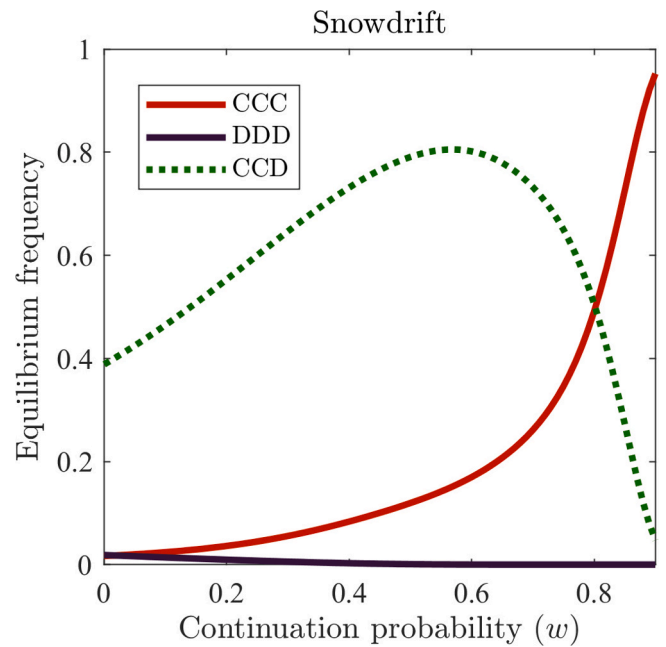


Fig. 5. Equilibrium frequency of strategies in the iterated SD. Equilibrium frequency of various strategies in a finite population playing the iterated SD, under a small-mutation approximation, as a function of the continuation probability (w). Continuous lines indicate power-independent strategies; the dotted line indicates (C, D, D), a strategy that is conditional on power differences. Of the 8 strategies studied, only the two ESSs and ALLD are shown. Other parameters: $\alpha = 0.5, b = 1, c = 0.4, N = 500, \beta = 0.05$.

on power changes the average level of cooperation in a finite population under a small-mutation approximation, calculated as the equilibrium frequency of a strategy times the fraction of interactions where that

strategy will play TFT (Fig. 6). These results confirm our claim that cooperation declines with power asymmetry (Fig. 6). This effect becomes more pronounced when the continuation probability is higher, and the decline in average cooperation increases with w . Thus, higher levels of power asymmetry cause a decline in the frequency of cooperative interactions.

4. Discussion

Power asymmetries are a pervasive feature of animal populations and human societies (Vallet et al., 2020; Scheffer et al., 2017; Piketty and Saez, 2014; Kaufmann, 1983; De Vries et al., 2006; Tibbetts et al., 2022). Modelling power asymmetry as the ability to allocate smaller or greater rewards, we investigate its impact on direct reciprocity in situations where power is either a stable or variable trait, showing that strong asymmetries can potentially destabilize reciprocal cooperation. In populations where power is a fixed trait, greater power asymmetry leads to a higher threshold continuation probability needed for stable cooperation to evolve and be stable (Equation (2), Fig. 1). In other words, for a certain frequency of repeated interaction, too much power asymmetry will eventually destabilize cooperation. An intuitive explanation for this phenomenon is that the cost-to-benefit ratio of cooperation for high-power individuals (who can allocate greater rewards, and receive smaller ones) increases with power asymmetry: if the level of asymmetry exceeds a critical threshold, the reward for mutual cooperation is so small that exploiting one's social partner once becomes more advantageous than sustained reciprocal cooperation. Therefore, it is in the interests of high-power individuals to defect, even when repeated interactions afford the possibility of reciprocal cooperation.

In symmetric interactions, heterogeneous social environments can select for the ability to infer interdependence (Colnaghi et al., 2023). Similarly, in social ecologies where power varies across interactions, there could exist adaptive benefits to inferring power differences and conditioning behavioural strategies of cooperation on power (Balliet et al., 2017; Kelley et al., 2003). Indeed, humans can readily detect, and respond to, differences in power (Gerpott et al., 2018; Redhead and Power, 2022; Hall et al., 2005; Fiske et al., 2007; Smith and Hofmann, 2016), and people employ a wide range of visual and auditory cues to infer differences in power (Carney, 2020; Aguinis et al., 1998; Carney et al., 2005). Children as young as five can accurately use nonverbal behaviour to discriminate between high- and low-power individuals (Brey and Shutts, 2015), and perceived power differences can, in turn, induce major changes in one's affective and cognitive state (Smith and Hofmann, 2016; Langner and Keltner, 2008; Berdahl and Martorana,

2006). Yet, not all social dilemmas facilitate the evolution of adaptations to infer differences in power.

Our work suggests that the ability to infer power differences across different interactions (symmetric, low-power, and high-power) is less likely to provide a strong selective advantage in interactions resembling the PD than in those resembling the Snowdrift. In the PD, power-independent strategies always outcompete strategies that are contingent on power asymmetries, except for a very narrow range of continuation probability (Fig. 3). By contrast, in the SD, a strategy that cooperates in symmetric and low-power interactions and defects in high-power interactions, (C, C, D) , is evolutionarily stable in infinite populations (Fig. 4) and favored by natural selection in finite populations (Fig. 5) for low to intermediate values of the continuation probability. Among the strategies that are conditional on power, (C, C, D) is the one that achieves the highest payoffs: when (C, C, D) interacts with itself in asymmetric interactions, the low-power individual cooperates, and the high-power individual therefore receives the highest payoff available to them. This strategy can outcompete (C, C, C) by exploiting it in low-power interactions, and (D, D, D) by cooperating with itself in symmetric interactions. Perhaps less obviously, (C, C, D) can also outcompete other strategies conditional on power differences, such as (C, D, C) . For example, a rare (C, C, D) mutant can invade a (C, D, C) population, but not vice versa, as the payoff of a low-power individual cooperating with a high-power individual is higher than that of a high-power individual cooperating with a low-power one.

This result suggests that the ability to infer differences in power provides a selective advantage in environments characterized by social dilemmas where anti-coordination is beneficial (i.e., when individuals can increase their payoffs by doing the opposite of what their partner does), such as the SD. In these games, the total payoffs are higher if individuals make opposite choices, especially because unilateral defection yields the best possible outcome whereas mutual defection yields the worst possible outcome. By contrast, when the continuation probability is high, the iterated PD becomes similar to a Stag Hunt (Skyrms, 2004); in this type of game, the total payoffs are higher if individuals make matching choices, especially because mutual cooperation yields the best possible outcome whereas unilateral cooperation yields the worst possible outcome.

It has been suggested that the emergence of power asymmetries, such as in the form of leadership and followership, provided a selective advantage in ancestral human societies, as it helped to solve collective action problems (King et al., 2009; Van Vugt et al., 2008; Pietraszewski, 2020). Repeated social interactions, in a context where a mixed group of leaders and followers performs better than a uniform group, give rise to

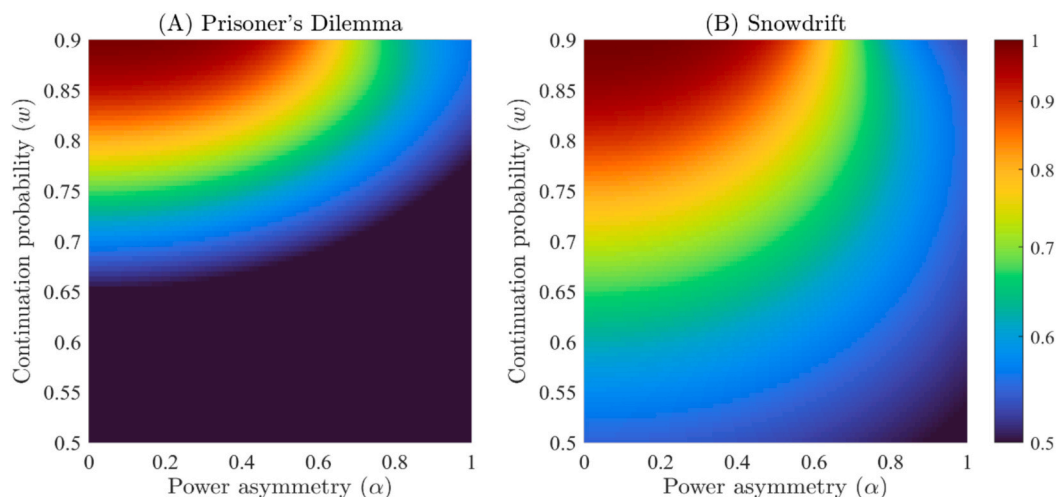


Fig. 6. Average frequency of cooperation. Average frequency of cooperation in a finite population under a small-mutation approximation, as a function of power asymmetry (α) and the continuation probability (w), in the iterated PD (A) and SD (B). Other parameters: $b = 1$, $c = 0.2$, $N = 50$, $\beta = 0.1$.

dynamics typical of the SD: if everyone strives to be a leader, an individual can increase their own (and collective) payoffs by being a follower; in a population of followers, becoming a leader will also lead to higher payoffs. Our model indicates that, under these conditions, the ability to infer differences in power helps maximize joint payoffs by facilitating anti-coordination, and the evolution of decision rules based on power can therefore be promoted by natural selection. We do not observe this in the iterated PD, where individuals can increase their payoffs by matching their partner's choice. In these games, strategies that discriminate based on power perform poorly against themselves, because they involve doing the opposite of what one's partner does when interactions are asymmetric.

Power asymmetries are often signalled by cues that are shared between people or other animals, and which can be used to facilitate coordination or anti-coordination (Hoffman et al., 2016). Social status is an example of cue that can be used to coordinate in social interactions. Prior research shows that people tend to defer to the preferences of higher-status individuals, even when status is arbitrarily assigned (Eckel and Wilson, 2007; Ball and Eckel, 1996; de Kwaadsteniet and van Dijk, 2010). This has been found to be particularly advantageous in asymmetric games, such as the battle of the sexes, where coordination would otherwise be difficult to achieve (de Kwaadsteniet and van Dijk, 2010). While social status and power are distinct constructs, they are closely associated, and cues of status can be used to infer power (Hall et al., 2005; Fiske et al., 2007).

Empirical studies on asymmetric games support the claim that asymmetry destabilizes cooperation (Hilbe et al., 2016; Ahn et al., 2007; Beckenkamp et al., 2006; Sheposh and Gallo, 1973; Talley, 1974). In asymmetric PDs, cooperation rates are significantly lower than in symmetric ones (Beckenkamp et al., 2006; Sheposh and Gallo, 1973). Moreover, information about the payoff matrix reduces cooperation rate in asymmetric games (Talley, 1974), while increasing it in symmetric ones (Talley, 1974; Gonzalez et al., 2015). Experimental studies also indicate that the temptation to defect is stronger for the players who have less to gain from cooperation (corresponding to the high-power players in our model) (Beckenkamp et al., 2006; Sheposh and Gallo, 1973; Talley, 1974). By contrast, in repeated asymmetric games, players who receive higher payoffs (corresponding to low-power individuals in our model) are more likely to initiate cooperation and less likely to defect in response to their partner's defection (Beckenkamp et al., 2006). These results indicate that, in agreement with our theoretical predictions, individuals who have less to gain from cooperation aim to minimize payoff disparity through frequent defection, thereby undermining the emergence of stable cooperation. While this behaviour is a suboptimal strategy in the PD (Fig. 3), it can provide a selective advantage in the asymmetric SD (Figs. 4 and 5).

Our results encourage empirical studies to investigate whether the tendency to defect when in the high-power state is stronger in asymmetric SD than PD games, supporting the hypothesis that this behaviour confers a selective advantage in games where anti-coordination is beneficial. While the studies discussed above indicate that payoff differences hinder cooperation, there is no universal consensus on the effects of power, and different sources of power asymmetry may yield distinct outcomes (Bone et al., 2016; Molho et al., 2019; Kopelman, 2009). For example, differences in the effectiveness of punishment promote cooperation in a modified version of the PD, where players can contribute a variable investment to the common good (Bone et al., 2016), but not in a standard PD, where players only have a binary choice to cooperate or defect (Nikiforakis et al., 2010). The asymmetric ability to punish is also more effective in promoting cooperation when used to encourage joint cooperation, rather than to exploit one's partner, in PD experiments (Kopelman, 2009). Different forms of asymmetry can also act synergistically with one another, as shown by a recent theoretical analysis of asymmetric public goods games (Hauser et al., 2019). This study shows that extreme power asymmetry, in the form of an unequal distributions of endowments, prevents the emergence of reciprocal

cooperation; however, when the rewards of cooperation are also asymmetric, unequal endowments actually promote, rather than hinder, cooperative behaviour (Hauser et al., 2019). Taken together, these studies suggest future theoretical and empirical work to further examine how different bases of power impact cooperation across different games and what factors could promote cooperation in the face of power asymmetry.

Some of our results are based on the assumption that individuals have an equal likelihood of finding themselves in a high- or low-power position. This assumption reflects the empirical observation that variations in people's perceptions of power are mainly due to changes in situations, rather than their stable traits (Smith and Hofmann, 2016). In addition, we used variations of the simultaneous donation game as a framework to study the emergence of reciprocal cooperation. In the SI, we discuss how this operationalisation of power is consistent with previous theoretical work on interdependence (68,69; Fig. S5). Including in our analysis both PD and SD games and varying the continuation probability allow us to explore a space of the four "archetypal" games most frequently studied in the literature on social dilemmas (Halevy et al., 2012; Santos et al., 2006; Peña and Nöldeke, 2023; Colnaghi et al., 2023). Changing the continuation probability transforms the evolutionary dynamics of the PD and the SD in interactions that resemble, respectively, the Stag Hunt and the Maximizing Difference games. Our choice of focussing on these four archetypal interactions is rooted in the historical work of Rapoport on 2x2 games (Rapoport, 1966) and reflected in previous theoretical research (Halevy et al., 2012; Santos et al., 2006; Peña and Nöldeke, 2023; Colnaghi et al., 2023). Yet, this is but a fraction of the 8-dimensional parameter space of all asymmetric, dyadic games. Further studies are needed to determine to what extent our conclusions can be generalized across all possible games.

Another important question concerns the exploration of alternative strategies of reciprocal cooperation. We only focused on TFT and ALLD: this is because under the assumptions of our model (no noise and no costs associated with TFT compared to other strategies), if TFT cannot outcompete ALLD, no other cooperative strategy can (Nowak, 2006). While TFT may pave the way for cooperation to evolve, it is not necessarily the best strategy to maintain cooperation once it has been established (Imhof et al., 2007; Nowak and Sigmund, 1993). In fact, TFT can be invaded by ALLC through drift, making it possible for ALLD to eventually take over the population (Imhof et al., 2007); on average, however, these evolutionary cycles tend to favour TFT (Imhof et al., 2005). Additionally, the benefits of TFT compared to ALLC increase with the continuation probability (Fig. S6), making TFT evolutionary advantageous precisely in those conditions (i.e., high w) that favour cooperation. As the main focus of this work is to establish the necessary conditions for cooperation to emerge, and TFT invading ALLD is the minimum requirement for cooperation to evolve (Nowak, 2006), we did not consider other strategies than TFT and ALLD (with the exception of Fig. S6, which also includes a strategy of unconditional cooperation, ALLC). Future investigations on cooperation in asymmetric games could explore a broader space of reactive and memory-one strategies, where one could find well-known strategies such as zero-determinant strategies (Taha and Ghoneim, 2020; Press and Dyson, 2012), win-stay lose-shift (Imhof et al., 2007; Nowak and Sigmund, 1993), or forgiving TFT (Godfray, 1992), and study the impact of alternative update rules in asymmetric games (Couto et al., 2022).

It is often said that "power corrupts", expressing the common wisdom that individuals in high-power positions can succumb to the temptation to exploit their subordinates (Pauwels et al., 2022). Yet, this is only one potential explanation why power asymmetries might undermine cooperation. Our results indicate that, when power is operationalized as the ability to provide higher rewards, power asymmetry makes it harder for reciprocal cooperation to evolve: high-power individuals have a stronger incentive to defect, because low-power individuals cannot provide enough benefits to their social partners to even out the costs of cooperation. However, conditioning one's decision to

cooperate on the level of power is not always an optimal strategy. In the iterated PD, it is more beneficial to adopt a power-independent strategy of either reciprocal cooperation (TFT) or defection (AllD), depending on the level of asymmetry and the continuation probability, regardless of power status (Figs. 2 and 3). In the iterated SD, on the other hand, power offers the means to maximize collective payoffs through anti-coordination: low-power individuals benefit from “carrying” the cost of cooperation, as they will receive the same treatment when they find themselves in a high-power position (Figs. 4 and 5). Regardless of whether the optimal strategy is conditional or not on the level of power, higher levels of asymmetry hinder the emergence of cooperation (Fig. 1) and thus lowers cooperation rates (Fig. 6). Power asymmetry thus undermines reciprocal cooperation, and interventions to enhance cooperative behaviour should aim at promoting a more egalitarian profitability of mutual cooperation.

CRedit authorship contribution statement

Marco Colnaghi: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Fernando P. Santos:** Writing – review & editing, Methodology, Conceptualization. **Paul A.M. Van Lange:** Writing – review & editing, Methodology, Conceptualization. **Daniel Balliet:** Writing – review & editing, Project administration, Methodology, Funding acquisition, Conceptualization.

Funding

This manuscript is part of a project that has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (Grant agreement No. 864519), awarded to Daniel Balliet.

Data availability

The code used to generate the results of this study has been deposited and is available on GitHub (<https://github.com/MColnaghi/power-cooperation>).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.jtbi.2025.112106>.

References

Van Lange, P.A.M., Balliet, D., Parks, C.D., Van Vugt, M., 2015. Social Dilemmas – The Psychology of Human Cooperation. Oxford University Press, Oxford.

Kollock, P., 1998. Social dilemmas: the anatomy of cooperation. *Annu. Rev. Sociol.* 24, 183–214.

Dawes, R.M., 1980. Social dilemmas. *Annu. Rev. Psychol.* 31 (1), 169–193.

Rand, D.G., Nowak, M.A., 2013. Human cooperation. *Trends Cogn. Sci.* 17 (8), 413–425.

Trivers, R.L., 1971. The evolution of reciprocal altruism. *Q. Rev. Biol.* 46, 35–57.

Axelrod, R., 1984. *The Evolution of Cooperation*. Basic Books, New York, NY.

Nowak, M.A., 2006. Five rules for the evolution of cooperation. *Science* 314 (5805), 1560–1563.

Schmid, L., Chatterjee, K., Hilbe, C., Nowak, M.A., 2021. A unified framework of direct and indirect reciprocity. *Nat. Hum. Behav.* 5 (10), 1292–1302.

Doebeli, M., Hauert, C., 2005. Models of cooperation based on the Prisoner’s Dilemma and the Snowdrift game. *Ecol. Lett.* 8 (7), 748–766.

Stewart, A.J., Plotkin, J.B., 2013. From extortion to generosity, evolution in the iterated Prisoner’s Dilemma. *PNAS* 110 (38), 15348–15353.

Imhof, L.A., Fudenberg, D., Nowak, M.A., 2007. Tit-for-tat or win-stay, lose-shift? *J. Theor. Biol.* 247 (3), 574–580.

Nowak, M.A., Sigmund, K., 1993. A strategy of win-stay, lose-shift that outperforms tit-for-tat in the Prisoner’s Dilemma game. *Nature* 1 (364), 56–58.

Van Veelen, M., Garcia, J., Rand, D.G., Nowak, M.A., 2012. Direct reciprocity in structured populations. *PNAS* 109 (25), 9929–9934.

Vallet, A., Locatelli, B., Barnaud, C., Makowski, D., Quispe Conde, Y., Levrel, H., 2020. Power asymmetries in social networks of ecosystem services governance. *Environ Sci Policy* 1 (114), 329–340.

Scheffer, M., Van Bavel, B., Van De Leemput, I.A., Van Nes, E.H., 2017. Inequality in nature and society. *PNAS* 114 (50), 13154–13157.

Piketty, T., Saez, E., 2014. Inequality in the long run. *Science* 344 (6186), 838–843.

Gerpott, F.H., Balliet, D., Columbus, S., Molho, C., de Vries, R.E., 2018. How do people think about interdependence? A multidimensional model of subjective outcome interdependence. *J. Pers. Soc. Psychol.* 115 (4), 716–742.

Redhead, D., Power, E.A., 2022. Social hierarchies and social networks in humans. *Phil Trans R Soc B.* 2022 Feb 28;377(1845):20200440.

Hall, J.A., Coats, E.J., LeBeau, L.S., 2005. Nonverbal behavior and the vertical dimension of social relations: a meta-analysis. *Psychol. Bull.* 131 (6), 898–924.

Fiske, S.T., Berdahl, J.L., 2007. Social power. In: Kruglanski, A.W., Tory Higgins, E. (Eds.), *Social Psychology: Handbook of Basic Principles*. The Guilford Press, New York, pp. 678–692.

Smith, P.K., Hofmann, W., 2016. Power in everyday life. *PNAS* 113 (36), 10043–10048.

Kaufmann, J.H., 1983. On the definitions and functions of dominance and territoriality. *Biol. Rev.* 58 (1), 1–20.

De Vries, H., Stevens, J.M.G., Vervaecke, H., 2006. Measuring and testing the steepness of dominance hierarchies. *Anim. Behav.* 71 (3), 585–592.

Tibbetts, E.A., Pardo-Sanchez, J., Weise, C., 2022. The establishment and maintenance of dominance hierarchies. *Phil Trans R Soc B.* 377, 1845.

Zimmaro, F., Miranda, M., Fernández, J.M.R., Moreno López, J.A., Reddel, M., Widler, V., et al., 2024. Emergence of cooperation in the one-shot Prisoner’s dilemma through Discriminatory and Samaritan AIs. *J. R. Soc. Interface* 21 (218), 20240212.

Han, T.A., Perret, C., Powers, S.T., 2021. When to (or not to) trust intelligent machines: insights from an evolutionary game theory analysis of trust in repeated games. *Cogn. Syst. Res.* 68, 111–124.

He, Z., Shen, C., Shi, L., Tanimoto, J., 2024. Impact of committed minorities: Unveiling critical mass of cooperation in the iterated prisoner’s dilemma game. *Phys. Rev. Res.* 6 (1), 013062.

Akata, Z., Balliet, D., De Rijke, M., Dignum, F., Dignum, V., Eiben, G., et al., 2020. A research agenda for hybrid intelligence: augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence. *Computer* 53 (8), 18–28.

Hammerstein, P.B., 1981. The role of asymmetry in animal contests. *Anim. Behav.* 29, 193–205.

Gaunersdorfer, A., Hofbauer, J., Sigmund, K., 1991. On the dynamics of asymmetric games. *Theor. Popul. Biol.* 39 (3), 345–357.

Maynard-Smith, J., Parker, G.A., 1976. The logic of asymmetric contest. *Anim. Behav.* 24, 159–175.

Dawkins, M.S., 2010. Do asymmetries destabilize the Prisoner’s Dilemma and make reciprocal altruism unlikely? *Anim. Behav.* 80 (2), 339–341.

Hauser, O.P., Hilbe, C., Chatterjee, K., Nowak, M.A., 2019. Social dilemmas among unequals. *Nature* 572 (7770), 524–527.

Ladret, V., Lessard, S., 2008. Evolutionary game dynamics in a finite asymmetric two-deme population and emergence of cooperation. *J. Theor. Biol.* 255 (1), 137–151.

Thibaut, J.W., Kelley, H.H., 1959. *The social psychology of groups*. New York: John Wiley.

Emerson, R.M., 1962. Power-dependence relations. *Am. Sociol. Rev.* 27, 31–41.

Ahn, T.K., Lee, M., Ruttan, L., Walker, J., 2007. Asymmetric payoffs in simultaneous and sequential prisoner’s dilemma games. *Public Choice* 132 (3–4), 353–366.

Beckenkamp, M., Frank, H.H.S., Maier-Rigaud, P., Hennig-Schmidt, H., Maier-Rigaud, P., Engel, C., et al., 2006. Cooperation in symmetric and asymmetric prisoner’s dilemma games. *MPI Collective Goods*. Preprint 25.

Sheposh, J.P., Gallo, P.S., 1973. Asymmetry of payoff structure and cooperative behavior in the prisoner’s dilemma game. *J. Confl. Resol.* 17 (2), 321–333.

Talley, M.B., 1974. Effects of asymmetry of payoff and asymmetry of information in a prisoner’s dilemma game. University of Texas, [Arlington].

Bone, J.E., Wallace, B., Bshary, R., Raihani, N.J., 2016. Power asymmetries and punishment in a prisoner’s dilemma with variable cooperative investment. *PLoS One* 11 (5).

Nikiforakis, N., Normann, H.T., Wallace, B., 2010. Asymmetric enforcement of cooperation in a social dilemma. *South. Econ. J.* 76 (3), 638–659.

Hilbe, C., Hagel, K., Milinski, M., 2016. Asymmetric power boosts extortion in an economic experiment. *PLoS One* 11 (10).

Molho, C., Balliet, D., Wu, J., 2019. Hierarchy, power, and strategies to promote cooperation in social dilemmas. *Games (Basel)* 10 (1).

Kopelman, S., 2009. The effect of culture and power on cooperation in commons dilemmas: Implications for global resource management. *Organ. Behav. Hum. Decis. Process.* 108 (1), 153–163.

Halevy, N., Chou, Y.E., Galinsky, D.A., 2011. A functional model of hierarchy. *Organ. Psychol. Rev.* 1 (1), 32–52.

Antonioni, A., Pereda, M., Cronin, K.A., Tomassini, M., Sánchez, A., 2018. Collaborative hierarchy maintains cooperation in asymmetric games. *Sci. Rep.* 8 (1), 5375.

French, J.R., Raven, B.H., 1959. The bases of social power. In: Cartwright, D. (Ed.), *Studies in Social Power*. University of Michigan, Ann Arbor, pp. 150–167.

Keltner, D., Gruenfeld, D.H., Anderson, C., 2003. Power, approach, and inhibition. *Psychol. Rev.* 110 (2), 265.

Maynard-Smith, J., 1978. The evolution of behavior. *Sci. Am.* 239 (3), 176–193.

- Bramoullé, Y., 2007. Anti-coordination and social interactions. *Games Econ Behav.* 58 (1), 30–49.
- King, A.J., Johnson, D.D.P., Van Vugt, M., 2009. The origins and evolution of leadership. *Curr. Biol.* 19 (19).
- Van Vugt, M., Hogan, R., Kaiser, R.B., 2008. Leadership, followership, and evolution: some lessons from the past. *Am. Psychol.* 63 (3), 182–196.
- Pietraszewski, D., 2020. The evolution of leadership: Leadership and followership as a solution to the problem of creating and executing successful coordination and cooperation enterprises. *Leadersh. Q.* 31 (2), 101299.
- Skyrms, B., 2004. *The stag hunt and the evolution of social structure*. Cambridge University Press.
- Halevy, N., Chou, E.Y., Murnighan, J.K., 2012. Mind games: the mental representation of conflict. *J. Pers. Soc. Psychol.* 102 (1), 132–148.
- Santos, F.C., Pacheco, J.M., Lenaerts, T., 2006. Evolutionary dynamics of social dilemmas in structured heterogeneous populations. *PNAS* 103, 3490–3494.
- Peña, J., Nöldeke, G., 2023. Cooperative dilemmas with binary actions and multiple players. *Dyn. Games Appl.* 13, 1156–1193.
- Colnaghi, M., Santos, F.P., Van Lange, P.A.M., Balliet, D., 2023. Adaptations to infer fitness interdependence promote the evolution of cooperation. *PNAS* 120 (50), e2312242120.
- Sigmund, K., 2010. *The Calculus of Selfishness*. Princeton University Press.
- Ohtsuki, H., 2010. Stochastic evolutionary dynamics of bimatrix games. *J. Theor. Biol.* 264 (1), 136–142.
- Nowak, M.A., Sasaki, A., Taylor, C., Fudenberg, D., 2004. Emergence of cooperation and evolutionary stability in finite populations. *Nature* 428 (6983), 646–650.
- Imhof, L.A., Fudenberg, D., Nowak, M.A., 2005. Evolutionary cycles of cooperation and defection. *PNAS* 102 (31), 10797–10800.
- Maynard-Smith, J., 1982. *Evolution and the Theory of Games*. Cambridge University Press, Cambridge.
- Fudenberg, D., Imhof, L.A., 2006. Imitation processes with small mutations. *J. Econ. Theory* 131 (1), 251–262.
- Moran, P.A.P., 1962. *The statistical process of evolutionary theory*. Clarendon Press, Oxford.
- Traulsen, A., Shores, N., Nowak, M.A., 2008. Analytical results for individual and group selection of any intensity. *Bull. Math. Biol.* 70 (5), 1410–1424.
- Couto, M.C., Giaimo, S., Hilbe, C., 2022. Introspection dynamics: a simple model of counterfactual learning in asymmetric games. *New J. Phys.* 24 (6), 063010.
- Balliet, D., Tybur, J.M., Van Lange, P.A.M., 2017. Functional interdependence theory: an evolutionary account of social situations. *Pers. Soc. Psychol. Rev.* 21 (4), 361–388.
- Kelley, H.H., Holmes, J.G., Kerr, N.L., Reis, H.T., Rusbult, C.E., Van Lange, P.A.M., 2003. *An Atlas of Interpersonal Situations*. Cambridge University Press, Cambridge.
- Carney, D.R., 2020. The nonverbal expression of power, status, and dominance. *Curr. Opin. Psychol.* 1 (33), 256–264.
- Aguinis, H., Simonsen, M.M., Pierce, C.A., 1998. Effects of nonverbal behavior on perceptions of power bases. *J. Soc. Psychol.* 138 (4), 455–469.
- Carney, D.R., Hall, J.A., LeBeau, L.S., 2005. Beliefs about the nonverbal expression of social power. *J. Nonverbal Behav.* 29 (2), 105–123.
- Brey, E., Shutts, K., 2015. Children use nonverbal cues to make inferences about social power. *Child Dev.* 86 (1), 276–286.
- Langner, C.A., Keltner, D., 2008. Social power and emotional experience: actor and partner effects within dyadic interactions. *J. Exp. Soc. Psychol.* 44 (3), 848–856.
- Berdahl, J.L., Martorana, P., 2006. Effects of power on emotion and expression during a controversial group discussion. *Eur. J. Soc. Psychol.* 36 (4), 497–509.
- Hoffman, M., Yoeli, E., Navarrete, C.D., 2016. Game theory and morality. *The Evolution of Morality*. 289–316.
- Eckel, C.C., Wilson, R.K., 2007. Social learning in coordination games: does status matter? *Exp. Econ.* 10 (3), 317–329.
- Ball, S.B., Eckel, C.C., 1996. Buying status: experimental evidence on status in negotiation. *Psychol. Mark.* 13 (4), 381–405.
- de Kwaadsteniet, E.W., van Dijk, E., 2010. Social status as a cue for tacit coordination. *J. Exp. Soc. Psychol.* 46 (3), 515–524.
- Gonzalez, C., Ben-Asher, N., Martin, J.M., Dutt, V., 2015. A cognitive model of dynamic cooperation with varied interdependency information. *Cogn. Sci.* 39 (3), 457–495.
- Rapoport, A., 1966. A taxonomy of 2×2 games. *General Systems*. 11, 203–204.
- Taha, M.A., Ghoneim, A., 2020. Zero-determinant strategies in repeated asymmetric games. *Appl. Math. Comput.* 15, 369.
- Press, W.H., Dyson, F.J., 2012. Iterated Prisoner's Dilemma contains strategies that dominate any evolutionary opponent. *PNAS* 109 (26), 10409–10413.
- Godfray, H.C.J., 1992. The evolution of forgiveness. *Nature* 16 (355), 206–207.
- Pauwels, L., Declerck, C.H., Boone, C., Diaz-Gutiérrez, P., Lambert, B., 2022. Does power corrupt? An fMRI study on the effect of power and social value orientation on inequity aversion. *J. Neurosci. Psychol. Econ.* 15 (4), 222–240.