Theory and Decision
https://doi.org/10.1007/s11238-021-09850-z

Classification by decomposition: a novel approach to classification of symmetric 2×2 games

Mikael Böörs[1] · Tobias Wängberg[2] · Tom Everitt[3,4] · Marcus Hutter[3,4]

Accepted: 10 October 2021

Abstract
In this paper, we provide a detailed review of previous classifications of 2×2 games and suggest a mathematically simple way to classify the symmetric 2×2 games based on a decomposition of the payoff matrix into a cooperative and a zero-sum part. We argue that differences in the interaction between the parts is what makes games interesting in different ways. Our claim is supported by evolutionary computer experiments and findings in previous literature. In addition, we provide a method for using a stereographic projection to create a compact 2-d representation of the game space.

Keywords Classification · Symmetric games · 2×2 Games · Decomposition · Cooperation and conflict · Simplicity

1 Introduction

What makes a game such as the Prisoner's Dilemma interesting? It is the tension between the common interest and self-interest, we argue in this paper. Indeed, in the Prisoner's Dilemma, the players' self interest directly opposes the common interest, often leading to mutual defection and tragedies of the commons. Meanwhile, in Stag Hunt, the conflict part pulls the players away from mutually more beneficial outcomes by counteracting the players' common interest. The Leader and Hero

Mikael Böörs and Tobias Wängberg contributed equally. The remaining authors contributed according to order.

✉ Tobias Wängberg
tobias@math.su.se

[1] University of Gothenburg, Gothenburg, Sweden

[2] Stockholm University, Stockholm, Sweden

[3] Google DeepMind, London, England, UK

[4] Australian National University, Canberra, Australia

Published online: 01 December 2021

games, also referred to as the symmetric Battle of the Sexes, exhibit yet another kind of tension, where the most payoff can be gained from alternating between outcomes in repeated games. Unfortunately, the conflict part makes such cooperation more difficult. Based on these observations we hypothesize that games with different tensions between common interest and self-interest should be interesting in different ways, and that differences in tensions is what distinguishes the standard games from each other. Indeed, all standard games (Prisoner's Dilemma, Chicken, Stag Hunt, Leader and Hero) exhibit different tensions between common interest and self-interest.

To explore this idea further, we establish a simple way to enumerate the possible tensions, and consider the classification of the space of 2×2 games that these *tension classes* induce (Sect. 4). Analysis of the regions show that they correctly separate the standard games, and make several further distinctions previously considered in the literature (Sect. 5). Computer experiments of iterated versions of the games give preliminary empirical support[1] for our hypothesis that tension classes separate the space of symmetric 2×2 games in strategically cohesive classes (Sects. 6 and 7).

A well-established scientific principle (e.g. Baker, 2007) says that a good scientific hypothesis should be:

1. simple and parsimonious, and
2. explain relevant observations.

Simplicity is important, as simple scientific theories have a much better track record, and tend to generalize better (cf. Occam's razor). In particular, classifications should avoid adding conditions ad hoc, as this increases complexity and tends to reduce generalisability. As for 2., a hypothesis that does not explain the relevant observations is either false or too vague. In the case of game classifications, a good classification hypothesis will separate significantly different games into different classes, while keeping similar games in the same class. Similarity between games has not been formally defined, but a consensus has emerged around which games should be grouped together and not (Harris, 1969; Huertas-Rosero, 2003; Rapoport et al., 1978).

Previous classifications have arguably failed to simultaneously satisfy both 1. and 2. Several works have suggested similar classes of games (Harris, 1969; Huertas-Rosero, 2003; Rapoport et al., 1978), but without basing the groupings on a simple principle. They therefore fail to meet 1. Other works have classified games based on simple mathematical principles (Borm, 1987; Robinson & Goforth, 2003), but with less convincing classes as a result, thereby failing to satisfy 2. Section 3 discusses these works in more detail.

In this paper, we argue that our classification based on the tension between common interest and conflict yields the right classes based on a simple principle, thereby satisfying both 1. and 2.

[1] One might argue that computer simulations of this kind do not count as empirical evidence. We do however hold the view that such simulation studies do provide some empirical support, but its significance for justifying our classification can be debated.

Outline. We begin with some background (Sect. 2), followed by a review of previous classifications of 2×2 games (Sect. 3). Focus is then shifted to our classification, defined in Sect. 4. The resulting regions are analyzed in Sect. 5. Computer experiments based on genetic algorithms provide preliminary empirical support for our classification (Sects. 6 and 7). We conclude the paper by summarizing our findings and presenting some open questions for further research (Sect. 8).

2 Preliminaries

In this section, we give a very brief review of game theory definitions. See for example the textbook by Gonzalez-Diaz et al. (2010) for more details and explanations. This section may be skipped by a reader knowledgeable in game theory.

Definition 2.1 (*Strategic game*) Let $\mathcal{P}$ be a set of players with $|\mathcal{P}| = n$. For all $i \in \mathcal{P}$, let S_i be the non-empty set of strategies of player i. Define the set of strategy profiles as $S \triangleq \prod_{i \in \mathcal{P}} S_i$. For all $i \in \mathcal{P}$, let $u_i : S \to \mathbb{R}$ be the pay-off function of player i and let $U \triangleq \{u_1, u_2, \ldots, u_n\}$ bet the set of payoff functions. A triple $G = (S, U, \mathcal{P})$ is called an *n-player strategic game* for the set of players $\mathcal{P}$.

Strategic games with 2 players are often represented as pairs of matrices. Each matrix represents the payoffs for each player. The position k, l in payoff matrix P_1 corresponds to the payoff player 1 receives from strategy profile $\{k, l\}$, i.e. $u_1(k, l) = P_{1_{k,l}}$. These payoff matrices are combined into a payoff bimatrix which represents the game. An example is shown in Fig. 1 where for example the strategy profile $s = \{0, 1\}$ results in payoff $u_1(0, 1) = b$ of player 1 and payoff $u_2(0, 1) = f$ of player 2.

Three important types of games.

We next define three special types of games that play a central role in our investigations: symmetric, zero-sum, and common-interest games.

Definition 2.2 (*Symmetric game*) A $n \times n$ game $G = (S, U, \mathcal{P})$ with $|\mathcal{P}| = 2$ and payoff bimatrix $P = (P_1, P_2)$ is called a *symmetric* game if

$$P_1 = P_2^{\mathrm{T}}.$$

In this paper we will focus on symmetric 2×2 games and often refer to the standard 2×2 symmetric game presented in Table 1.

$$P_1 = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \qquad P_2 = \begin{bmatrix} e & f \\ g & h \end{bmatrix} \qquad P = (P_1, P_2) = \begin{bmatrix} (a, e) & (b, f) \\ (c, g) & (d, h) \end{bmatrix}$$

Fig. 1 The payoff matrices of player 1 and player 2 and the resulting payoff bimatrix

Table 1 Standard symmetric 2×2 game

		Player 2	
		0	1
Player 1	0	(a, a)	(b, c)
	1	(c, b)	(d, d)

Both players can choose between action 0 and action 1. The outcomes of the games are denoted ij where $i \in \{0, 1\}$ is the action of player 1 and $j \in \{0, 1\}$ is the action of player 2

A zero-sum game is a game where the payoffs of every strategy profile add up to zero. Intuitively this means that what one player gains, the other player loses. Hence the two players have completely opposite interests in this type of game.

Definition 2.3 (*Zero-sum game*) An $m \times n$ game $G = (S, U, \mathcal{P})$ with $|\mathcal{P}| = 2$ and payoff bimatrix $P = (P_1, P_2)$ is called a *zero-sum game* if

$$P_1 + P_2 = 0_{m \times n}.$$

The opposite of a zero-sum game is a common interest game. In a common interest game, the two players get the same payoff in every outcome and therefore have the same interests.

Definition 2.4 (*Common interest game*) An $m \times n$ game $G = (S, U, \mathcal{P})$ with $|\mathcal{P}| = 2$ and payoff bimatrix $P = (P_1, P_2)$ is called a *common interest game* if

$$P_1 = P_2.$$

Strategic equivalence. To reduce the number of cases, it is natural to consider games strategically equivalent if their payoffs are positive linear transformations of each other, or if their players or actions are renamed. Under reasonable assumptions, such translations never affect the strategic analysis of the game.

Definition 2.5 (*Strategic equivalence*) Two $n \times n$ games with payoff bimatrices P and P' respectively are said to be *strategically equivalent*, denoted $P \sim P'$, if P' can be obtained by permuting the rows, columns or players in P (renaming players or actions), or if

$$\exists a \in \mathbb{R}, b \in \mathbb{R}_+ : P' = aJ_n + bP$$

where J_n is the $n \times n$ matrix of ones (positive linear transformation of payoffs).

Equilibria concepts. Nash equilibrium, abbreviated NE, is an important concept in game theory. Intuitively if a strategy profile is a NE, it means that no player has an incentive to change strategy given that no other player changes strategy.

Definition 2.6 (*Nash equilibrium*) Given a game $G = (S, U, \mathcal{P})$, a strategy profile $s \in S$ is said to be a *Nash equilibrium* if

$$\forall i \in \mathcal{P} u_i(s_{-i}, s_i) \geqslant u_i(s_{-i}, s_i'), \quad \forall s_i' : s_i' \neq s_i.$$

A concept related to NE is the concept of Altruistic Equilibrium, abbreviated AE. It was used by Huertas-Rosero (2003) (although under a different name), to classify symmetric 2×2 games. AE is not as commonly known as NE, but as we will see in this paper it can help us understand some aspects of many games. Intuitively, a strategy profile is AE if neither player has an incentive to change his strategy given that he tries to maximize his opponents' payoff. In 2×2 games, this means that AE is equivalent to NE in the transposed payoff bimatrix.

Definition 2.7 (*Altruistic equilibrium*) Given a game with a set of strategy profiles S and a set of payoff functions U, a strategy profile $s \in S$ is said to be an *Altruistic Equilibrium* if

$$\forall i \in \mathcal{P} \forall j \in \mathcal{P} : j \neq i \; u_j(s_{-i}, s_i) \geqslant u_j(s_{-i}, s_i'), \quad \forall s_i' : s_i' \neq s_i.$$

In the literature review, we will consider dominant strategies and best response correspondences.

Definition 2.8 (*Strictly dominant strategy*) A strategy s_i for player i is called *strictly dominant* if

$$\forall s_i' : s_i' \neq s_i \quad \forall s_{-i} u_i(s_{-i}, s_i) > u_i(s_{-i}, s_i').$$

Intuitively Definition 2.8 states that a strategy is dominant if every other strategy yields a lower payoff, regardless of what strategy the other players use.

Definition 2.9 (*Dominant strategy equilibrium*) A game is said to have a *dominant strategy equilibrium* if there exists a strategy profile $s \in S$ such that all strategies in s are dominant.

If the dominant strategy equilibrium exists, it is unique, this follows immediately from its definition.

Definition 2.10 (*Best response correspondence*) Let $G = (S, U, \mathcal{P})$ be a strategic game such that for all $i \in \mathcal{P}$ there exists $n_i \in \mathbb{N}$ such that $S_i \subset \mathbb{R}^{n_i}$, $S_i \neq \emptyset$ and S_i is compact. The correspondence $B_i : S_{-i} \to S_i$ is called the *best response correspondence* for player i and is defined as

$$B_i(s_{-i}) \triangleq \{s_i^* \in S_i : u_i(s_{-i}, s_i^*) \geq u_i(s_{-i}, s_i') \forall s_i' \in S_i\}.$$

for any given $s_{-i} \in S_{-i}$

If a strategy profile $s^* \in S$ is such that $s_i^* \in B_i(s_{-i}^*)$ for every player $i \in \mathcal{P}$, then s^* is a Nash equilibrium. Definition 2.10 can be extended to mixed extension games and this extension is used in Sect. 3.3.

2.1 Standard games

Below we list the symmetric 2×2 games, referred to as the symmetric standard games. For more information about these games, see for example Rapoport et al. (1978). The Leader and Hero games are also referred to as symmetric versions of the asymmetric standard game Battle of the Sexes (Harris, 1969).

	0 (Cooperate)	1 (Defect)
0 (Cooperate)	$(3,3)$	$(1,4)$
1 (Defect)	$(4,1)$	$(2,2)$

Prisoner's Dilemma

	0 (Stag)	1 (Hare)
0 (Stag)	$(3,3)$	$(0,2)$
1 (Hare)	$(2,0)$	$(1,1)$

Stag Hunt

	0 (Continue)	1 (Give up)
0 (Continue)	$(-2,-2)$	$(1,-1)$
1 (Give up)	$(-1,1)$	$(0,0)$

Chicken

	0 (Sacrifice)	1 (Exploit)
0 (Sacrifice)	$(1,1)$	$(3,5)$
1 (Exploit)	$(5,3)$	$(2,2)$

Hero

	0 (Exploit)	1 (Sacrifice)
0 (Exploit)	$(2,2)$	$(3,5)$
1 (Sacrifice)	$(5,3)$	$(1,1)$

Leader

3 Literature review

In this section, we review five different approaches to classifying 2×2 games. The focus of the review is to provide an accessible explanation of each work, and to evaluate how well they satisfy the following desiderata:

1. Simplicity and parsimony, and
2. Well-justified regions.

As discussed in the introduction, simplicity is important for generalization. Well-justified regions mean that the classification groups games in a way that corresponds to important strategic considerations. One would for example expect that the well established standard games discussed in Sect. 2.1 are considered different types of games in a classification. The primary conclusion of the review is that none of the reviewed works satisfy both desiderata. A more detailed review can be found in Böörs and Wängberg (2017), Chapter 2.2.

3.1 A hierarchical approach to classifying 2×2 games (Rapoport et al., 1978)

In the book *The* 2×2 *Game*, Rapoport et al. (1978) present a taxonomy of all 2×2 games (not just the symmetric ones). The authors create a discretized version of the game space by considering games within an ordinal scale. Each player has 4 payoffs, ranked from lowest to highest, that can be placed in 4 different cells. The game space is therefore made up of $4! \times 4! = 576$ games. By grouping strategically equivalent games (Sect. 2.5), they reduce the number of classes to 78, which they organize hierarchically according to five different properties. Inspired by systems of classification in biology, the authors divide the game space first into different phyla, then divide each phylum into classes, followed by orders and finally genera. This system of classifying is perhaps the most intuitive and a natural approach to classifying objects of any kind. It does not require any advanced theory and a structured division of the game space is gained.

The authors use several interesting properties to sort the games. They use concepts such as degree of conflict, which means how aligned the players' interests are. For example, the game has no conflict if both players' interests are aligned on the same outcome. In contrast, a zero-sum game is a game with full conflict. They also classify by the existence of different kinds of *pressure*. Pressures are used to describe different situations where the players may have an incentive to deviate from the Nash equilibrium. Depending on the type of conflict and the amount of pressure acting on the game, the authors label the games as having different degrees of stability. The concept of pressure becomes relevant in repeated play, where the players might be tempted (or forced) to change from their equilibrium strategies. An example is competitive pressure, which is described as a game where a player prefers an outcome where he gets a higher relative payoff compared to the opposing player, and would thereby deviate from an equilibrium outcome even though it would result in a lower absolute payoff.

Rapoport et al. provide extensive experimental support for their classification, establishing the relevancy of their classification to human strategic play in 2×2 games. This classification therefore fulfills Desideratum 2. Although a structured and systematic sorting of the game space is achieved, the classification lacks in mathematical simplicity. It therefore fails to satisfy Desideratum 1.

3.2 A parameter based classification of 2×2 symmetric games (Harris, 1969)

Harris (1969) classifies the symmetric[2] 2×2 games by positioning them on a plane according to the following two parameters:

[2] Harris (1969) classifies a generalization of symmetric games that he calls interval-symmetric. An interval-symmetric game is a game that is symmetric up to a positive linear transformation of one of the players' payoff matrix.

$$r_3 = \frac{d-b}{c-b}$$
$$r_4 = \frac{c-a}{c-b} \quad (1)$$

where a, b, c and d are the players' payoffs, as in Table 1. The constraint that $c > b$ is also imposed. This is to make the inequalities for r_3 and r_4 in terms of the payoff parameters a, b, c and d unambiguous, and means that player 1 gets smaller payoff for outcome $\{0, 1\}$ and larger for $\{1, 0\}$.

Harris introduces inequalities on the parameters that partition the (r_3, r_4)-plane into 12 different regions based on the signs and relative size of r_3 and r_4. An advantage of this approach is its mathematical structure and simplicity. Since any desired inequality of the payoffs can be expressed using the r_3 and r_4 parameters, this classification is easily modified to other classification conditions. One can also algebraically compare it to other classifications because of the mathematical structure. Furthermore, the parametrisation approach enables a visual representation of the game space. Unfortunately, Harris method lacks a clear principle for selecting inequalities by which to partition the plane. Therefore, it does not satisfy Desideratum 1, which we identify as a weakness of this classification. On the other hand, the author puts a lot of effort into justifying the resulting classes, and we therefore conclude that Desideratum 2 is fulfilled.

3.3 A classification of 2×2 bimatrix games (Borm, 1987)

In the article *A classification of* 2×2 *bimatrix games*, Borm (1987) defines a classification that partitions the space of mixed 2×2 games into 15 classes. The main idea of the classification is to divide games into different classes based on the structure of the two players' best reply correspondences (BRC), which is an extension of Definition 2.10 to mixed extension games. Borm shows that for the mixed extension of 2×2 games there are four basic types of BRCs for each player (Fig. 2). The backbone in the classification is unordered pairs (one for each player) of these BRC types. There are in 10 such unordered pairs in total.

The pair of the players' BRCs in a game is important because it has a close connection to Nash equilibria, as a mixed strategy profile is NE if and only if it is in the intersection of the two players BRC. However, even though there are 10 BRC combinations there are 15 types of NE sets in mixed 2×2 games. Borm argues that it is important to isolate all 15 types of NE in different classes, and as Fig. 3 illustrates, the BRC pairs alone do not. Therefore Borm defines four variables as linear combinations of the payoffs. The combination of the BRC pairs and the value of the four variables are enough to isolate all 15 types of NE in different classes and these 15 classes are the final partition of the space of mixed 2×2 games i Borm's classification.

This method of classification differs from the ones previously presented in several ways. It is, for example, a classification system that applies to mixed 2×2 games, not only strictly ordered or pure ones. It also differs in that it classifies games based on fewer concepts than the others, essentially just taking Nash

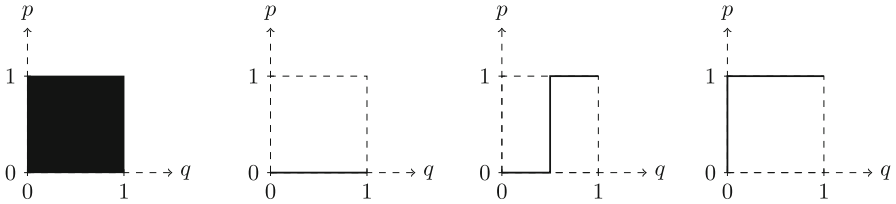
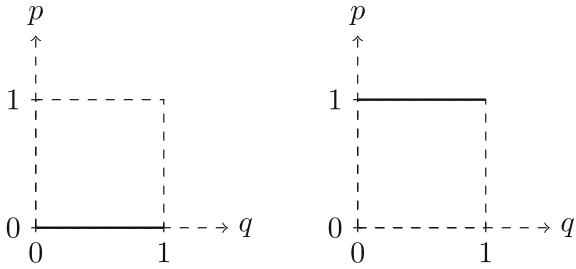


Fig. 2 Suppose that both players can choose to play action a_1 or action a_2. If *player 1* chooses a_1 with probability p and *player 2* chooses a_1 with probability q then every mixed strategy profile is a point (p, q) in the unit rectangle. For every q we can plot the BRC of *player 1* and since there are four basic types of BRC there are four basic types of the resulting graph. These are the four graphs illustrated in the figure. The graph of *player 2's* BRC can be illustrated as the transpose of one of the graphs in the figure

Fig. 3 Graphs illustrating two NE sets with the same BRC type



equilibria into account. This might be considered a disadvantage, since this means the classification misses some important distinctions, for example the alignment of the common interest and the self-interests of the players. This (dis-)alignment accounts for the difference between vastly different types of games, such as Prisoner's Dilemma and Deadlock which have the same NE but differs in this alignment. However, the simplicity of this method is also a strength since it allows Borm to divide the set of all mixed 2×2 games into just 15 classes, which is a relatively small number. Despite this we conclude that the classification is simple and parsimonious, i.e. that Desideratum 1. is fulfilled. Unfortunately, Borm provides no empirical or theoretical evidence that suggests that all of his classes are interestingly different. We conclude that the classification fails to fulfil Desideratum 2.

3.4 A cartography for 2×2 symmetric games (Huertas-Rosero, 2003)

In the article *A Cartography for* 2×2 Symmetric Games, Huertas-Rosero (2003) presents a classification that divides the space of symmetric 2×2 non-zero sum games into 12 different classes[3] based on their type of Nash equilibria and on their type of Altruistic Equilibria (see Definition 2.7). Huertas-Rosero observes that a symmetric 2×2 non-zero sum game can have NE in either one diagonal outcome, in both diagonal outcomes or both anti-diagonal outcomes and that the same is true

[3] We have proved that one of Huertas-Rosero's 12 classes is empty (Böörs & Wängberg, 2017, chapter 2.2.4).

for AE. The 8 combinations of the type of NE and the type of AE forms the foundation for this classification.

Huertas-Rosero has a rather elegant way of defining his 8 base classes. By defining an isometry that allows him to express the game space in well-chosen parameters in $\mathbb{R}^4$, he can express the NE and AE conditions using only two parameters each. Using the properties of additive- and multiplicative invariance he fixates one parameter to 0 and normalize the resulting game vector so that every game is represented as a point on the unit sphere in $\mathbb{R}^3$. The NE and AE conditions defines four pairwise orthogonal planes through the origin that split the sphere into 14 pieces. Some of the 14 regions are strategically equivalent, and after taking this into account Huertas obtains his 8 base classes. The 8 base classes are then divided further into 12 classes based on whether the NE payoffs are higher than the AE payoffs or not.

Huertas-Rosero's geometrical approach to representing and classifying games is mathematically simple. The NE and AE conditions used as a base for the classification are easily expressed within the geometrical representation, and the conditions for the classification could be generalized to higher dimensions. In addition, Huertas-Rosero does not introduce any ad hoc conditions to divide any of his classes. We therefore conclude that the classification satisfies Desideratum 1. The concept of NE is commonly accepted as an important concept in game theory and therefore we believe that Huertas-Rosero is justified to use this as a base for his classification. However, the concept of AE is not well known and Huertas-Rosero does not justify why it is an important concept. Neither does the author justify why the AE condition defines interestingly different regions. Hence the classification fails to satisfy Desideratum 2.

3.5 A topologically-based classification of the 2×2 ordinal games (Robinson & Goforth, 2003)

In the article *A Topologically-Based Classification of the* 2×2 *Ordinal Games* by Robinson, a classification system based on defining a topology on the 2×2 games is proposed. Robinson criticise the hierarchical classification system of Rapoport et al. presented in Sect. 3.1. They claim that a lot of games with no topological relation can end up in the same class when classifying in this way. The authors formalize what it means for games to be similar to each other, which they use to create the topology. They argue that using the hierarchical approach, the deep relations between games are often not understood. The authors therefore propose a topological approach to classifying all 2×2 games which they claim captures how games are related to each other.

In general, a topology is a set of points together with a neighbourhood relation. In this case, the points will be the classes of 2×2 games with the same ordinal relationship between the payoffs. The neighborhood relation is based on the smallest possible change you can make to a game that affects strategic play. When changing the payoff function continuously, strategic play changes only when the preference order is changed. The authors therefore define the operation, which we will denote

$S_{i,p}(G)$, that swaps the i:th ranked payoff for the $i+1$:th of player p in game G. For example, using $S_{1,2}$ on the game in Table 2 results in the game in Table 3 where the lowest and second-lowest payoff for player 2 have been swapped. Following the reasoning above leads to Definition 3.1.

Definition 3.1 (*Neighbouring games*) The *set of neighbours* $N(G)$ of game G is defined as

$$N(G) = \{S_{i,p}(G) : i \in \{1,2,3\}, p \in \{1,2\}\}.$$

The definition states that a game G is a neighbour of G' if $S_{i,p}(G) = G'$ for some $i \in \{1,2,3\}$, $p \in \{1,2\}$. Note that every game has a total of 6 neighbours since there are 6 possible swap operators.

The topological space is created by starting with an arbitrary game and then applying all possible concatenations of the $S_{i,p}$ operation on that game. This results in a space containing 144 unique 2×2 games. Depending on which game is chosen to start with and what swap operations are used, different configurations of the map will be created. By restricting to fewer swap operations than the 6 defined above you can create *subspaces* of the topological space. The classification is made by partitioning the topological space into a number of closed subspaces, containing games that are topologically similar to each other. Depending on what subspaces are investigated, different classifications will be obtained. The authors, for example, begin by investigating the subspace created by restricting to the $S_{1,p}$ and $S_{2,p}$ swaps. They motivate this by that swapping the lowest-ranked payoffs likely has a lower impact on strategic play than swapping the highest-ranked payoff. The games contained in this subspace would therefore be similar to each other.

The topological approach is in our view an elegant way of representing the game space. The advantage of this approach is therefore clearly its mathematical structure, i.e. that it fulfils Desideratum 1. However, even though the topological approach has some justification, the resulting classes lack further empirical or theoretical justification. Hence we conclude that Desideratum 2 is not fulfilled.

Table 2 Payoff matrix before using the $S_{1,2}$ swap operation

		Player 2	
		0	1
Player 1	0	(2, 4)	(4, 3)
	1	(1, 2)	(3, 1)

Table 3 Payoff matrix after using the $S_{1,2}$ swap operation

		Player 2	
		0	1
Player 1	0	(2, 4)	(4, 3)
	1	(1, 1)	(3, 2)

3.6 Conclusions

The conclusion from this literature review is that none of the classification approaches here reviewed are able to fulfill both Desiderata 1. and 2. The conclusions are summarized in Table 4.

4 Classification by decomposition

In this section, we introduce our classification. Section 4.1 presents a theorem stating that any 2×2 game can be decomposed into a common interest game and a zero-sum game, and show how this can be used to describe the tension between the common interest and the self interest of the players. Section 4.2 then considers the classification of symmetric 2×2 games this gives rise to.

4.1 Decomposition of symmetric 2×2 games

Any game can be decomposed into a common interest game and a zero-sum game (Kalai & Ehud, 2013). This is shown in Proposition 4.1 using the following natural convention for addition of bimatrices:

Table 4 This table summarises which of the desiderata defined in the beginning of this section that are satisfied

Author	Desideratum 1: simplicity and parsimony	Desideratum 2: well-justified regions
Rapoport et al. (1978)	No	Yes
Harris (1969)	No	Yes
Borm (1987)	Yes	No
Huertas-Rosero (2003)	Yes	No
Robinson and Goforth (2003)	Yes	No

Each classification receives a Yes if it manages to fulfil a desideratum and a No otherwise

$$(P_1, P_2) + (P'_1, P'_2) = (P_1 + P'_1, P_2 + P'_2).$$

Proposition 4.1 (Decomposition) *A 2-player game with payoff bimatrix P can always be decomposed into the sum of a zero-sum game with payoff bimatrix Z and a game of common interest with payoff bimatrix C:*

$$P = C + Z.$$

Proof Let P be the payoff bimatrix of an arbitrary 2-player game. P can be decomposed as follows:

$$\begin{aligned} P &= (P_1, P_2) \\ &= \frac{1}{2}((P_1 + P_2), (P_1 + P_2)) + \frac{1}{2}((P_1 - P_2), (P_2 - P_1)) \\ &= C + Z. \end{aligned}$$

Here $C = \frac{1}{2}((P_1 + P_2), (P_1 + P_2))$ is a common interest game (Definition 2.4), and $Z = \frac{1}{2}((P_1 - P_2), (P_2 - P_1))$ is a zero-sum game (Definition 2.3). □

Theorem 4.1 shows how a general 2-player game bimatrix can be decomposed into one common interest game and a zero-sum game. We will interchangeably refer to the zero-sum game as the conflict part of the game, since the players' incentives are completely opposed in these games. In the special case of 2×2 symmetric games, the decomposition of the payoff matrix has the following form:

$$\begin{aligned} \begin{bmatrix} (a,a) & (b,c) \\ (c,b) & (d,d) \end{bmatrix} &= \begin{bmatrix} (a,a) & (\frac{b+c}{2}, \frac{b+c}{2}) \\ (\frac{b+c}{2}, \frac{b+c}{2}) & (d,d) \end{bmatrix} \\ &+ \begin{bmatrix} (0,0) & (\frac{b-c}{2}, -\frac{b-c}{2}) \\ (-\frac{b-c}{2}, \frac{b-c}{2}) & (0,0) \end{bmatrix}. \end{aligned} \tag{2}$$

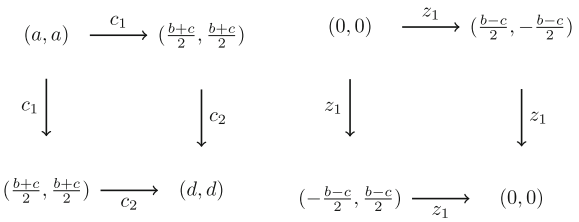


Fig. 4 The payoffs to the left is the common interest part of the payoff matrix and to the right is the conflict part of the payoff matrix. In the picture the eight variables that describe the preferences of both players are illustrated as arrows

The decomposition separates the common interest part from the conflict part of the game. We define the variables c_1, c_2 and z_1 for the incentives in the common interest and the zero-sum parts (see Fig. 4). In the definition, u_i^C and u_i^Z are the payoffs for player i in the common interest part and the zero-sum part, respectively.

$$c_1 = u_1^C(10) - u_1^C(00) = \frac{b+c}{2} - a \tag{3}$$

$$c_2 = u_2^C(11) - u_2^C(10) = d - \frac{b+c}{2} \tag{4}$$

$$z_1 = u_1^Z(10) - u_1^Z(00) = -\frac{b-c}{2} - 0 = -\frac{b-c}{2}. \tag{5}$$

***Remark* 4.2** In symmetric 2×2 games we only need c_1, c_2 and z_1 to describe the incentives in the common interest and the zero-sum parts. For general 2×2 games one could also define $c_3 = u_2^C(01) - u_2^C(00), c_4 = u_1^C(11) - u_1^C(01)$ and define z_2, z_3 and z_4 analogously. However in symmetric 2×2 games $c_1 = c_3, c_2 = c_4$ and $z_k = z_l \forall k, l \in \{1, 2, 3, 4\}$ as shown in Fig. 4.

The variables c_1, c_2 and z_1 allow us to compare the strength of the incentives in the common interest part with the strengths of the incentives in the conflict part, and to check whether the incentives are aligned or not. For example, if $c_1 > 0$, then player 1 prefers outcome $\{1, 0\}$ in the common interest matrix over outcome $\{0, 0\}$. The sign of the variables determine what outcomes the players prefer. In Fig. 4 we represent these preferences as arrows, where a positive sign of the variables mean that the arrows point in the direction shown in the figure, and in the opposite direction if the sign is negative. The vertical arrows represent the preferences of player 1 and the horizontal arrows represent the preferences of player 2. The absolute value of the variables c_1, c_2 and z_1 represent the strengths of the players' preferences.

In the five standard games (Prisoner's Dilemma, Chicken, Stag Hunt, Leader and Hero), the conflict part of the decomposition counteracts the common interest part of the decomposition. Expressed in our variables, this means that the conflict variable z_1 and at least one of the common interest variables c_1 or c_2 have opposite

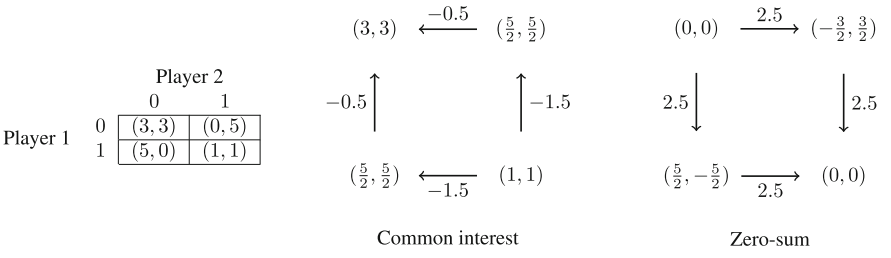


Fig. 5 The common interest part (middle) and the conflict part (right) of the decomposed payoff bimatrix of the Prisoner's Dilemma payoff matrix (left)

signs and that the value of z_1 is greater than this variable. Consider for example the decomposition of Prisoner's Dilemma in Example 4.3.

***Example* 4.3** In Prisoner's Dilemma, the players have to choose between cooperation and defection. The temptation of the selfish choice is stronger than the incentive to cooperate and as result $\{1, 1\}$ (both players defect) is the unique NE even though both players could get a higher payoff from cooperating. The game below is an example of a Prisoner's Dilemma game. In Fig. 5 below, the decomposition of the Prisoner's Dilemma bimatrix is presented. As suggested by the arrows, the conflict part of the decomposition draws the players toward the selfish outcome and it is strong enough to counteract both of the common interest variables that draw the players toward the cooperation outcome.

In the common interest game both players prefer outcome $\{0, 0\}$ (cooperation) and hence the arrows point in this direction. In the conflict game player 1 prefers the $\{1, 0\}$ outcome and player 2 prefers the $\{0, 1\}$ outcome. As a result, their combined incentives draw them toward the $\{1, 1\}$ outcome, as the arrows suggest. Because the payoff differences in the conflict game is larger than those in the common interest game, the incentives in the conflict game counteracts the incentives in the common interest game. This can be told from the conflict and common interest variables. We have $z_1 = 1.5$ and $c_1 = c_2 = -0.5$. Because the z_1 variable and the c_1 and c_2 variables have different signs the conflict counteracts the common interest and since $|z_1| > \max(|c_1|, |c_2|)$, the conflict is strong enough to overpower the common interest. Therefore the selfish outcome $\{1, 1\}$ is the NE of the game.

Strategic equivalence Games with different payoffs and tension vectors may be strategically equivalent. Indeed, except for the zero-conflict game with $z_1 = 0$, all games are equivalent to a game with $z_1 = 1$.

Proposition 4.4 *Every game with non-zero conflict is strategically equivalent with a game with* $z_1 = 1$.

Proof According to Definition 2.5, if the payoff matrix of a game can be constructed by permuting the rows and columns of another game's payoff matrix, then the games are strategically equivalent. This means that the two matrices in Eq. 6 are strategically equivalent:

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \sim \begin{bmatrix} d & c \\ b & a \end{bmatrix} = A'. \tag{6}$$

These matrices have the following decomposition.

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} a & \frac{b+c}{2} \\ \frac{b+c}{2} & d \end{bmatrix} + \begin{bmatrix} 0 & \frac{b-c}{2} \\ -\frac{b-c}{2} & 0 \end{bmatrix} \tag{7}$$

$$A' = \begin{bmatrix} d & c \\ b & a \end{bmatrix} = \begin{bmatrix} d & \frac{b+c}{2} \\ \frac{b+c}{2} & a \end{bmatrix} + \begin{bmatrix} 0 & -\frac{b-c}{2} \\ \frac{b-c}{2} & 0 \end{bmatrix}. \tag{8}$$

That is, $c_1 = -c_2'$, $c_2 = -c_1'$ and $z_1 = -z_1'$ and from this we can tell that every point in $\mathbb{R}^3$ with $z_1 < 0$ represents a game that is strategically equivalent with a game such that $z_1 > 0$. Hence we only need to consider games with $z_1 \geq 0$.

Scalar invariance defined in Definition 2.5 states that to games with payoff matrices A and A' respectively are strategically equivalent if there exists $\alpha > 0$ such that $A' = \alpha A$. Multiplying the game matrix with a constant α is equivalent with multiplying all of the payoffs with α. Since $z_1 = -\frac{b-c}{2}$, the relation $A' = \alpha A$ implies that $z_1' = -\frac{\alpha b - \alpha c}{2} = \alpha z_1$. This means that every game with $z_1 > 0$ has a strategically equivalent game with $z_1 = 1$. Combined with the observation above, this completes the proof. □

4.2 Classification

The common interest and conflict variables c_1, c_2, z_1 locate games in $\mathbb{R}^3$ via the vector $[c_1, c_2, z_1]^t$. It is easily shown that every game with negative z_1 is strategically equivalent with a game with positive z_1 (see proof of Proposition 4.4). A partition of $\mathbb{R}^2 \times \mathbb{R}_+$ thereby groups games into classes. The arguably simplest partitioning of $\mathbb{R}^2 \times \mathbb{R}_+$ is given by the planes spanned by any two coordinate axes:

$$c_1 = 0 \tag{9}$$

$$c_2 = 0. \tag{10}$$

This turns out to be a far too coarse classification, with significantly different games such as Prisoner's Dilemma and Stag Hunt sharing regions. The second simplest set of planes is arguably the diagonal planes:

$$c_1 = c_2 \quad c_1 = -c_2 \tag{11}$$

$$c_1 = z_1 \quad c_1 = -z_1 \tag{12}$$

$$c_2 = z_1 \quad c_2 = -z_1. \tag{13}$$

Splitting $\mathbb{R}^2 \times \mathbb{R}_+$ with the planes defined in Eqs. 9–13 gives a accurate game classification.

Proposition 4.5 *The planes* $c_1 = 0, c_2 = 0, z_1 = \pm c_1, z_1 = \pm c_2$ *and* $c_1 = \pm c_2$ *divide* $\mathbb{R}^2 \times \mathbb{R}_+$ *into* 24 *different regions.*

Proof The first two planes defines 2 regions each and since they are pairwise orthogonal this results in $2^2 = 4$ regions in total. The lengths $|c_1|, |c_2|$ and $|z_1|$ can be

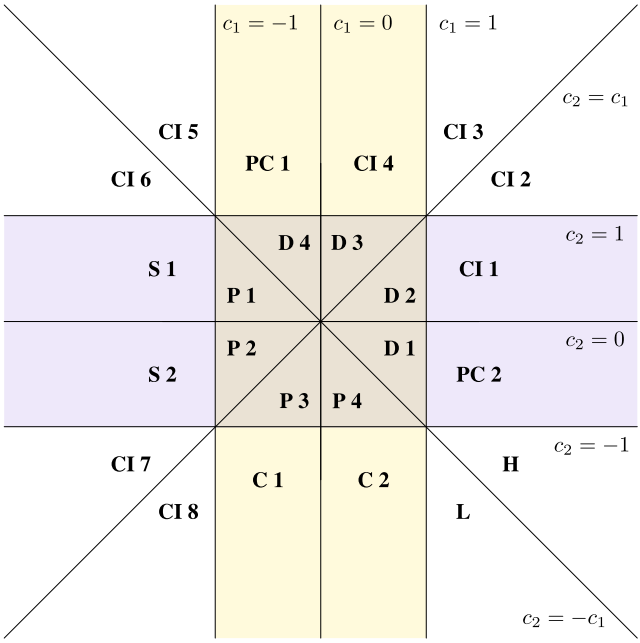


Fig. 6 An illustration of the partition of the space of symmetric 2×2 games into 24 classes. The lines are defined by the intersection of the $z_1 = 1$ plane and the planes described in Proposition 4.5. In the yellow area $|z_1| > |c_1|$, the conflict incentive is stronger than the cooperation incentive c_1. Similarly, in the blue area $|z_1| > |c_2|$, the conflict incentive is stronger than the cooperation incentive c_2. In the middle rectangle where z_1 dominates both c_1 and c_2, the conflict is stronger than the common interest incentives. When z_1 and c_k have the same sign, $k \in \{1, 2\}$, the incentives z_1 and c_k point in the same direction. Since $z_1 = 1$ in the map, when c_k is positive there is no tension between the interests represented by z_1 and c_k. Therefore the first quadrant is a no-conflict area. In the outer white regions, $|z_1| < \min(|c_1|, |c_2|)$, i.e. the common interest is stronger than the conflict. Most of the standard games are found in the more central, coloured regions, where the conflict is stronger (colour figure online)

ordered in $3! = 6$ ways independently of the signs of c_1 and c_2. Hence the planes divide $\mathbb{R}^3$ into $2^2 \times 3! = 24$ regions in total. □

By employing Proposition 4.4 we can visualise the classification through its intersection with $z_1 = 1$. The visualisation is shown in Fig. 6.

As we will establish in the next subsection, all of the standard games have their own regions. For example, the four Prisoner's Dilemma regions can be found in the high-conflict part of the map. The *Common Interest* regions consist of games where the conflict variable is weaker than both of the common interest variables, or where the conflict variable and the common interest variables have the same sign. In the *Partial Conflict* regions the conflict variable and one of the common interest

variables have opposite signs and the conflict variable is strong enough to dominate the common interest variable.

4.3 Interpreting the boundaries

What explains the success of this classification principle is that the planes generated by Eqs. 9–13 capture points where either the alignment between common interest and conflict turns to disalignment, or the relative strength between two components c_1, c_2, z_1 shifts. In the following subsections we analyse interpretations and implications of the planes more closely, and link our 24 regions to the some of the standard games. We also determine where Nash equilibria (NE), Definition 2.6, and Altruistic Equilibria (AE), Definition 2.7, are located in different regions.

4.3.1 Bijective transformation

As a first step for our more careful analysis, we will add an auxiliary variable x in addition to c_1, c_2 and z_1, to make the transformation from the payoff parameters a, b, c, d bijective. In Fig. 4 we define the variables c_1, c_2 and z_1 according to the transformation

$$\begin{bmatrix} c_1 \\ c_2 \\ z_1 \end{bmatrix} = \frac{1}{2} \begin{bmatrix} -2 & 1 & 1 & 0 \\ 0 & -1 & -1 & 2 \\ 0 & -1 & 1 & 0 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix}. \tag{14}$$

The transformation in Eq. 14 is not invertible. To make it invertible, we add the auxiliary variable $x \triangleq \frac{a+b+c+d}{2}$. The resulting transformation

$$\begin{bmatrix} x \\ c_1 \\ c_2 \\ z_1 \end{bmatrix} = A \begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix} = \frac{1}{2} \begin{bmatrix} 1 & 1 & 1 & 1 \\ -2 & 1 & 1 & 0 \\ 0 & -1 & -1 & 2 \\ 0 & -1 & 1 & 0 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix} \tag{15}$$

is invertible, with determinant $|A| = -1$.

Note that changing the auxiliary variable x is equivalent to adding a constant to the payoffs. It preserves strategic equivalence. This in combination with the fact that the transformation in Eq. 15 is invertible and Proposition 4.4 implies that every point in the map in Fig. 6 represents exactly one strategic equivalence class of games. It also means that every strategic equivalence class of games is represented by exactly one point in the map. This is desirable since there is no interesting difference between strategically equivalent games. Hence it is not meaningful to represent the same strategically equivalence class with more than one point in a classification map.

By inverting the transformation matrix A in Eq. 15, we can express the payoffs a, b, c, d, in terms of c_1, c_2, z_1 and x:

$$\begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix} = A^{-1} \begin{bmatrix} x \\ c_1 \\ c_2 \\ z_1 \end{bmatrix} = \frac{1}{4} \begin{bmatrix} 2 & -3 & -1 & 0 \\ 2 & 1 & -1 & -4 \\ 2 & 1 & -1 & 4 \\ 2 & 1 & 3 & 0 \end{bmatrix} \begin{bmatrix} x \\ c_1 \\ c_2 \\ z_1 \end{bmatrix}. \tag{16}$$

Being able to convert between our variables c_1, c_2, z_1 and the payoffs a, b, c, d will be useful in establishing Propositions 4.7 and 4.8 below. The next three subsections will investigate each of the planes generated by Eqs. 9–13 in turn.

4.3.2 The $c_i = \pm z_1$ conditions, and NE and AE regions

We next state and prove two propositions expressing conditions for NE and AE in terms of our cooperation and conflict variables c_1, c_2 and z_1. The concept of NE is standard in game theory. The concept of AE is not as commonly known as NE, but it does nevertheless play an important part in many games, for example *Deadlock*, *Chicken* and in the symmetric version of *Battle of the Sexes*.

To be able to express the NE and AE conditions in terms of c_1, c_2 and z_1 we need to introduce Lemma 4.6.

Lemma 4.6 *The following list states the NE conditions expressed in terms of a, b, c and d.*

1. $\{0,0\}$ *is NE iff* $a \geq c$.
2. $\{0,1\}$ *is NE iff* $b \geq d$ *and* $c \geq a$.
3. $\{1,0\}$ *is NE iff* $c \geq a$ *and* $b \geq d$.
4. $\{1,1\}$ *is NE iff* $d \geq b$.

Proof We provide the proof for the first case. The proof for the other three cases are analogous. According to Definition 2.6, $\{0,0\}$ is NE if and only if $u_1(0,0) \geqslant u_1(1,0)$ and $u_2(0,0) \geqslant u_2(0,1)$. But $u_1(0,0) \geqslant u_1(1,0) \Leftrightarrow a \geq c$ and $u_2(0,0) \geqslant u_2(0,1) \Leftrightarrow a \geq c$. That is $\{0,0\}$ is NE iff $a \geq c$. □

Proposition 4.7 *In any symmetric 2×2, two-player game G with cooperation and conflict variables c_1, c_2 and z_1, the following statements are true*:

(i) $\{0,0\}$ *is NE if and only if* $z_1 + c_1 \leq 0$,
(ii) $\{0,1\}$ *and* $\{1,0\}$ *are NE if and only if* $z_1 + c_1 \geq 0$ *and* $z_1 + c_2 \leq 0$,
(iii) $\{1,1\}$ *is NE if and only if* $z_1 + c_2 \geq 0$.

Proof Let $\vec{y} = (a, b, c, d)^{\mathrm{T}}$ be a vector with the payoffs for G, and let

$$\vec{v} = (x, c_1, c_2, z_1)^{\mathrm{T}} = A\vec{y},$$

where A is the matrix in Eq. 15. Equation 16 gives us that the payoff vector $\vec{y}$ can be written $\vec{y} = A^{-1}\vec{v}$, and hence

$$a = \frac{1}{4}(2x - 3c_1 - c_2) \quad b = \frac{1}{4}(2x + c_1 - c_2 - 4z_1)$$
$$c = \frac{1}{4}(2x + c_1 - c_2 + 4z_1) \quad d = \frac{1}{4}(2x + c_1 + 3c_2). \tag{17}$$

Combining the list and Eq. 17 with Lemma 4.6, we can express these conditions in terms of the variables c_1, c_2 and z_1.

(i) $\{0,0\}$ is NE if and only if

$$a \geq c \iff \frac{1}{4}(2x - 3c_1 - c_2) \geq \frac{1}{4}(2x + c_1 - c_2 + 4z_1) \iff z_1 + c_1 \leq 0.$$

(ii) $\{0,1\}$ and $\{1,0\}$ are NE if and only if $b \geq d$ and $c \geq a$.

$$b \geq d \iff \frac{1}{4}(2x + c_1 - c_2 - 4z_1) \geq \frac{1}{4}(2x + c_1 + 3c_2) \iff z_1 + c_2 \leq 0.$$

Similarly, $c \geq a \iff z_1 + c_1 \geq 0$.

(iii) $\{1,1\}$ is NE if and only if

$$d \geq b \iff z_1 + c_2 \geq 0.$$

□

Proposition 4.7 implies that the planes $z_1 + c_1 = 0$ and $z_1 + c_2 = 0$ divide $\mathbb{R}^3$ into four regions with different types of NE. Both of the planes that separates games with different types of NE are used in our model to partition the game space. This in every region in our model, all of the games have the same type of NE. Recall that in Fig. 6, $z_1 = 1$, which means that the lines $c_1 = -1$ and $c_2 = -1$ divide the map into the four NE regions (see Fig. 7).

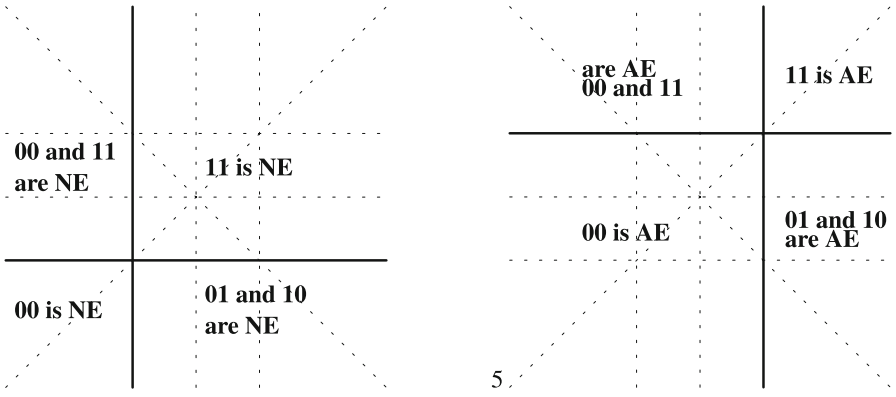


Fig. 7 In the left map the bold lines are the NE conditions from Proposition 4.7 and they divide the plane into four regions with different types of NE. In the right map the bold lines are the AE conditions from Proposition 4.8 and they divide the plane into four regions with different types of AE. It is easy to see that there are nine possible combinations of NE and AE in symmetric 2×2 games

Proposition 4.8 provides the analogous analysis for AE. The lines $c_1 = 1$ and $c_2 = 1$ divide the games according to their AE types in the same way that $c_1 = -1$ and $c_2 = -1$ divide the NE regions.

Proposition 4.8 *In any symmetric* 2×2, *two-player game G with cooperation and conflict variables* c_1, c_2 *and* z_1, *the following statements are true*:

(i) $\{0,0\}$ *is AE iff* $z_1 \geq c_1$,
(ii) $\{0,1\}$ *and* $\{1,0\}$ *are AE iff* $z_1 \leq c_1$ *and* $z_1 \geq c_2$,
(iii) $\{1,1\}$ *is AE iff* $z_1 \leq c_2$.

Proof Let $\vec{y} = (a,b,c,d)^{\mathrm{T}}$ be a vector with the payoffs for G, and let

$$\vec{v} = (x, c_1, c_2, z_1)^{\mathrm{T}} = A\vec{y},$$

where A is the matrix in Eq. 15. We can express the payoffs a , b, c and d associated with $\vec{v}$ as in Eq. 17. In a 2×2 symmetric game AE is equivalent to NE in the transposed payoff matrix. This means that

1. $\{0,0\}$ is AE iff $a \geq b$,
2. $\{0,1\}$ and $\{1,0\}$ are AE iff $b \geq a$ and $c \geq d$,
3. $\{1,1\}$ is AE iff $d \geq c$.

Using Eq. 17, these conditions can be expressed in the variables c_1, c_2 and z_1 as follows.

(i) $\{0,0\}$ is AE if and only if

$$a \geq b \iff \frac{\alpha}{4}(2x - 3c_1 - c_2) \geq \frac{\alpha}{4}(2x + c_1 - c_2 - 4z_1) \iff z_1 \geq c_1.$$

(ii) $\{0,1\}$ and $\{1,0\}$ is AE if and only if $b \geq a$ and $c \geq d$.

$$b \geq a \iff z_1 \leq c_1$$

$$c \geq d \iff \frac{\alpha}{4}(2x + c_1 - c_2 + 4z_1) \geq \frac{\alpha}{4}(2x + c_1 + 3c_2) \iff z_1 \geq c_2.$$

(iii) $\{1,1\}$ is AE if and only if

$$d \geq c \iff z_1 \leq c_2.$$

□

Proposition 4.8 states that all four types of AE are described by the planes $z_1 - c_1 = 0$ and $z_1 - c_2 = 0$. Just as the NE planes, they divide $\mathbb{R}^3$ into four different regions with different types of AE and the planes are used in our model to describe the strength relationship between $|z_1|$, $|c_1|$ and $|c_2|$. In Fig. 7 the four NE regions and the four AE regions in the map from Fig. 6 are shown.

Propositions 4.7 and 4.8 show that our method of classification, which is intended to divide games into groups depending on their type of decomposition,

captures all different types of NE and AE. Since NE and AE[4] have been argued to be important aspects of games, the results support our hypothesis that tension between the common interest and self-interest is what makes a game interesting, and that games are interesting in different ways because they have different common interest–self-interest tensions.

4.3.3 The $c_1 = \pm c_2$ conditions

Now we know the meaning of the planes $z_1 = \pm c_1, c_2$. The next step is to analyze the meaning of the planes $c_1 = \pm c_2$. The plane $c_1 = -c_2$ determines the difference between the two diagonal outcomes, as

$$c_1 > -c_2 \iff -2a + b + c > b + c - 2d \iff a > d.$$

Thus, the players prefer outcome $\{0, 0\}$ over $\{1, 1\}$ when $c_1 > -c_2$, and the $\{1, 1\}$ outcome over the $\{0, 0\}$ outcome otherwise. An interpretation of this condition is that when $c_1 + c_2 > 0$, the common interests are more aligned than disaligned with the zero-sum incentive.

This difference is exactly the difference between *Prisoner's Dilemma* and *Deadlock*. In Prisoner's Dilemma $c_1 + c_2 < 0$, meaning that the summed common interests are disaligned with the zero-sum incentive. In Deadlock on the other hand $c_1 + c_2 > 0$ and hence the summed common interests are aligned with the zero-sum incentive. Indeed, in Fig. 6 the $c_1 = -c_2$ line is the border between Prisoner's Dilemma and Deadlock.

The plane $c_1 = c_2$ is not a border between any standard games. It does however have some interesting properties. Note that

$$c_1 > c_2 \iff -2a + b + c > -b - c + 2d \iff b + c > a + d.$$

This means that when $c_1 > c_2$, the mean of the payoffs in the anti-diagonal positions are higher than the mean of the diagonal positions. Therefore, in some of the regions $c_1 > c_2$, the players can cooperate by alternating between the two anti-diagonal positions in iterated play. However the $c_1 + c_2 = 0$ condition is arguably not as important as the $c_1 - c_2 = 0$ condition. Remember that the latter decides whether or not the common interest incentives are more aligned than disaligned with the zero sum incentives. The $c_1 - c_2 = 0$ condition has no direct connection to the zero sum incentives and therefore one might expect this condition to have a smaller impact on the games than the $c_1 + c_2 = 0$ condition.

4.3.4 The $c_i = 0$ conditions

The $c_i = 0$ conditions decide the common interest incentives and in general the direction of the c_i arrows are important. However, since the c_i incentives are small near the boarder $c_i = 0$ they will be dominated by the zero sum incentives.

[4] We do not in fact see the strategic relevance of AE, but mention this to stress that our classification do capture information about this. We found that the AE separating planes have less impact on strategic play compared to the NE separating planes for example.

Therefore it seems likely that the change of direction of the c_i arrows will not make a clear distinction between interestingly different games. Indeed, in the map in Fig. 6 these condition does not separate any standard games. Instead one might expect a gradual change in the games as c_i changes signs.

4.4 Comparison with other classifications

In this section we compare our proposed classification with those reviewed in Sect. 3 that have a similar approach as us. We prove that our classification is at least as fine-grained as the classifications by Harris and Huertas-Rosero. Moreover, we capture the conditions which Harris add ad hoc to divide certain regions further. Since our approach to classifying games differs quite a bit from the classifications by Rapoport et al. and Robinson and Goforth, we do not compare our classification with theirs. The classification by Borm is also hard to compare with, since he classifies mixed extension 2×2 games. We leave these comparisons as open questions in Sect. 8.

4.4.1 Huertas-Rosero (2003)

Huertas-Rosero (2003) classifies non-zero sum symmetric 2×2 games based on NE and AE locations, and on whether or not the payoffs in the NE outcomes are larger than the payoffs in the AE outcomes. As stated in Propositions 4.7 and 4.8, the NE and AE locations are determined by our classes. Further, in each of our 24 classes one can tell whether the payoffs in the NE outcomes are larger than the payoffs in the AE outcomes. A proof for this can be found in Böörs and Wängberg (2017), Chapter 3.2. This means that all of Huertas-Rosero's classes except one can

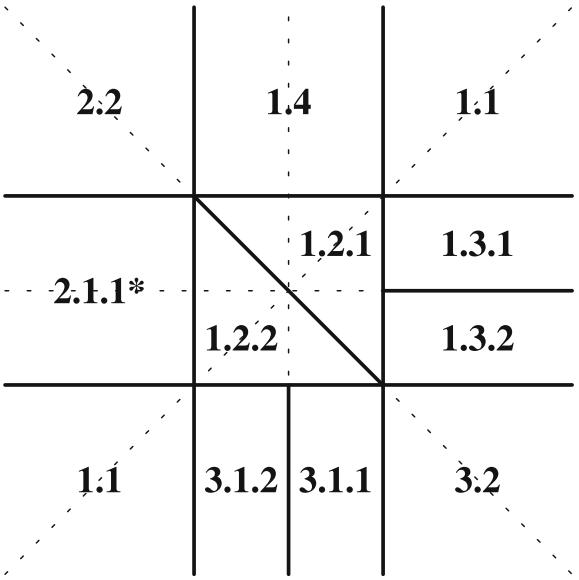


Fig. 8 The 11 non-empty classes of Huertas-Rosero named as in Huertas-Rosero (2004). The twelfth and empty class is the part of 2.1.1 that lies above the line $c_2 = -c_1$. As suggested by the figure, Huertas-Rosero fails to capture the different sub-regions of Prisoner's Dilemma and Stag Hunt, and the two standard games Leader and Hero belong to the same class in his classification

be found in our classification. However, this remaining class is empty (Böörs and Wängberg 2017, chapter 2.2.4). Thus, our classification contains all 11 non-empty classes of Huertas-Rosero. Figure 8 shows Huertas-Rosero's 11 non-empty classes in our classification map.

All the aspects of the classification provided by Huertas-Rosero are found in our classification as well, despite his rather different starting point based on NE and AE locations. This supports our hypothesis that the interesting differences between games are explained by the differences in their decomposition.

4.4.2 Harris (1969)

Harris uses a geometrical approach to classifying games (see Sect. 3.2). In that respect his method is similar to our classification method. We can therefore easily compare the parameters that are used in Harris1969' classification to ours, and prove that they are in fact equivalent. Harris defines the two parameters

$$r_3 = \frac{d-b}{c-b} \quad r_4 = \frac{c-a}{c-b}.$$

He partitions the space of symmetric 2×2 games by dividing the $r_3 r_4-$plane with the lines defined by

$$r_3 = 1 \quad r_4 = 0 \quad r_3 = 0 \quad r_3 + r_4 = 1 \quad r_4 = 1. \tag{18}$$

Supposing that $c \neq b$, we can express these lines in terms of our variables using the definition of r_3 and r_4 and Eq. 17. If $c = b$, the game lies on the border between two of our classes:

$$\begin{aligned} r_3 = 1 &\iff c = d \iff z_1 = c_2 \\ r_4 = 0 &\iff c = a \iff z_1 = -c_1 \\ r_3 = 0 &\iff d = b \iff z_1 = -c_2 \\ r_4 = 1 &\iff b = a \iff z_1 = c_1 \\ r_3 + r_4 = 1 &\iff d = a \iff c_2 = -c_1. \end{aligned} \tag{19}$$

Harris also uses two ad hoc conditions to divide only his Prisoner's Dilemma region. These extra conditions also correspond to conditions in our classification, as can be easily demonstrated:

$$\begin{aligned} r_4 = \frac{1}{2} &\iff b + c = 2a \iff c_1 = 0 \\ r_3 = \frac{1}{2} &\iff b + c = 2d \iff c_2 = 0. \end{aligned} \tag{20}$$

In Fig. 9 Harris' classes are drawn in our map to show the similarities and the differences between his classification and ours. Harris' classes are a subset to the set of our classes. This again supports the idea that the interesting elements of a game is defined by its decomposition. Our classification also captures the different variants of Prisoner's Dilemma and Chicken without the need for ad hoc conditions.

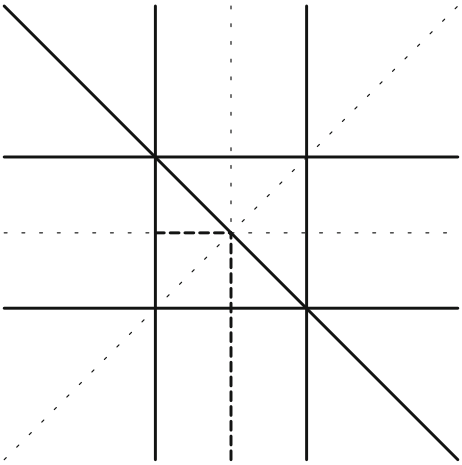


Fig. 9 The classes of Harris. The 12 main classes that are the result of the lines in Eq. 19 are drawn with bold lines. As can be seen in the map, all standard games are isolated in Harris1969' classification. With the extra conditions of Eq. 20 depicted with bold dashed lines, he divides the Prisoner's Dilemma and the Chicken regions into sub-regions to capture the different variations of the games. However, he does not capture the sub-regions in Stag Hunt and he does not differentiate between the non-conflict games CI 1 and 4 and the partial conflict games PC 1 and 2 (cf. Fig. 6)

5 Analysis of game regions

In this section we analyse the regions associated with standard games, and compare them with relevant literature. The comparisons are made both with the classifications reviewed in Sect. 3, and other research. Many of our regions have attracted interest from authors approaching the topic from very different starting points. In the analysis we will be referring to the payoff parameters a, b, c and d, as in Table 5.

5.1 Prisoner's dilemma regions

In one-shot scenarios where the game is played only once the variations are not very interesting, as $\{1, 1\}$ is the only NE (see Table 6). However, the variations become interesting in iterated play, where a player may sacrifice short term payoff to increase the chances of the opponent cooperating in the future (Axelrod & Hamilton, 1981).

Table 5 Symmetric 2×2 game with payoff parameters a, b, c and d

	0	1
0	(a, a)	(b, c)
1	(c, b)	(d, d)

Table 6 Four representatives of the P1, P2, P3 and P4 regions

	0 (Cooperate)	1 (Defect)
0 (Cooperate)	$(7,7)$	$(0,8)$
1 (Defect)	$(8,0)$	$(5,5)$

P1

	0 (Cooperate)	1 (Defect)
0 (Cooperate)	$(3,3)$	$(-1,5)$
1 (Defect)	$(5,-1)$	$(0,0)$

P2

	0 (Cooperate)	1 (Defect)
0 (Cooperate)	$(4,4)$	$(-1,5)$
1 (Defect)	$(5,-1)$	$(1,1)$

P3

	0 (Cooperate)	1 (Defect)
0 (Cooperate)	$(2,2)$	$(0,6)$
1 (Defect)	$(6,0)$	$(1,1)$

P4

Harris (1969) distinguishes between P1 and P4 and regards P2 and P3 as the same game. All the versions of PD above are regarded as the same game by Huertas-Rosero (2003), Rapoport et al. (1978) and Robinson and Goforth (2003) reviewed in Sect. 3

In terms of the payoff parameters, the Prisoner's Dilemma regions correspond to:

P1 $b + c < 2d$
P2 $2d < b + c < 2a$ and $a - c < b - d$
P3 $2d < b + c < 2a$ and $a - c > b - d$
P4 $b + c > 2a$.

In P4, the players may start to alternate between the strategy profiles $\{0,1\}$ and $\{1,0\}$ because it gives both players higher payoff than cooperating when the game is repeated. Some argue that this makes the distinction between cooperation and defection too vague, since the players can cooperate by taking turns exploiting each other (Axelrod & Hamilton, 1981; Harris, 1969).

In the P1 region, the inequality $b + c < 2d \iff b - d < d - c$ holds. This means that the cost for signalling for cooperation $(b - d)$ is less in magnitude than the gain of inducing the opposing player to initiate cooperation first $(d - c)$. Some argue that this is not a true representative of the Prisoner's Dilemma (Harris, 1969; Lave, 1965).

Therefore many researchers refer to the *Restricted* Prisoner's Dilemma shown in Table 6 (P2 and P3), where the inequality $2d < b + c < 2a$ is satisfied, as the true version of the game (Axelrod & Hamilton, 1981; Harris, 1969; Radinsky, 1971; Rapoport & Albert, 1965; Scodel, 1962). Many also use representations of PD where this inequality is satisfied, without stating it (Andreoni & Miller, 1993; Lave, 1965).

The Restricted Prisoner's Dilemma is however divided into two different games, depending on whether $c_1 > c_2$ (P2) or whether $c_1 < c_2$ (P3).

We found no previous discussion of the P2–P3 distinction. We did however find an intuitive meaning: it affects the cost of signalling for cooperation and the cost for accepting it given that the game is repeated and in NE.

For example, in the P3 instance in Table 6, signalling is expensive compared to accepting the signal to cooperate. The cost for signalling, by switching from action

1 to action 0, and "telling the opposing player that you wish to cooperate" is $|b-d| = |-1-1| = 2$, but the cost of responding to the signal is only $|a-c| = |4-5| = 1$. In the P2 example in Table 6 the situation is reversed. The cost for signalling is lower whereas the cost for accepting the signal is comparatively high. In this case player 1 pays a small cost of $|d-b| = |0-1| = 1$ to signal to player 2 that he would like to cooperate and player 2 pays a higher cost of $|a-c| = |3-5| = 2$ for agreeing to cooperate. We leave it as an open question whether this has any significant impact on strategic play.

5.2 Stag Hunt regions

The Stag Hunt regions are characterised by having two NE, with one of the Nash equilibria Pareto dominating all other outcomes. What makes the Stag Hunt games interesting is a conflict incentive that although weak, may still push cautious, non-trusting players towards the worse NE. The first version of Stag Hunt S1 has a weaker push towards good NE with $c_2 > 0$, whereas the second one S2 has a stronger push towards the good NE with $c_2 < 0$. Examples of the games are displayed in Table 7.

Stag Hunt was overlooked in the classification by Harris (1969) and called a no-conflict game. However, this game has received a lot of attention from other authors (Dubois et al., 2012; Rapoport et al., 1978; Skyrms, 2004).

A Nash equilibrium outcome is said to be payoff-dominant if it is not strictly Pareto dominated by any other outcome. A low-risk equilibrium can be interpreted as the intersection of the maximin strategies introduced by Rapoport et al. (1978). Several authors have analyzed Stag Hunt to see whether players will choose the payoff-dominant or the low-risk equilibrium strategies (Dubois et al., 2012; Rapoport et al., 1978).

A natural conjecture is that $c_2 < 0$ (S2) makes the payoff-dominant equilibrium more frequent and that $c_2 > 0$ (S1) makes the risk-dominant equilibrium played more often. Dubois et al. (2012) conducted human experiments where three different versions of Stag Hunt, displayed in Table 8, were played repeatedly over 75 rounds. The authors found that the frequency of cooperation (payoff-dominant strategy) was higher in Game 1 (S2), with $c_2 = -9 < 0$ than in Game 2 (S1) and

Table 7 Two variants of Stag Hunt, one from each region

	0 (Stag)	1 (Hare)
0 (Stag)	$(4,4)$	$(-1,3)$
1 (Hare)	$(3,-1)$	$(2,2)$

Stag Hunt 1 (S1)

	0 (Stag)	1 (Hare)
0 (Stag)	$(5,5)$	$(0,4)$
1 (Hare)	$(4,0)$	$(1,1)$

Stag Hunt 2 (S2)

Cooperation tends to be more frequent in the right game (Dubois et al. 2012), which may be explained by the common interest incentive c_2 being positive in S1 and negative in S2

Table 8 The three versions of Stag Hunt used in the experiments by Dubois et al. (2012) together with parameter values

	0	1
0	$(45, 45)$	$(0, 42)$
1	$(42, 0)$	$(12, 12)$

Game 1 (S2) $c_2 = -9$

	0	1
0	$(40, 40)$	$(20, 37)$
1	$(37, 20)$	$(32, 32)$

Game 2 (S1) $c_2 = 3.5$

	0	1
0	$(44, 44)$	$(4, 38)$
1	$(38, 4)$	$(28, 28)$

Game 3 (S1) $c_2 = 7$

A result from the experiments was that the frequency of cooperation over the last five rounds (71–75) was 43.75% in game 1, 24.38% in game 2 and 30.94% in game 3

Table 9 Cooperation dominant game

	0	1
0	(6, 6)	(3, 5)
1	(5, 3)	(2, 2)

Game 3 (S1) with $c_2 = 3.5 > 0$ and $c_2 = 7 > 0$ respectively. The authors also found that cooperation was slightly more frequent in Game 3 than in Game 2, but that the difference became negligible in the last 25 rounds. This provides support for the conjecture above, that there is an important difference between region S1 and S2.

Rapoport et al. (1978) discuss the difference between Stag Hunt and the Cooperation Dominant game found in regions CI7 and CI8 (see Table 9). Although similar, the difference is that $c_2 < -z_1$ in the Cooperation Dominant game, which makes cooperation a dominant strategy. Unsurprisingly, empirical results by Rapoport et al. showed that cooperation is much more frequent in the Cooperation Dominant game than in Stag Hunt.

5.3 Chicken regions

Similar to Stag Hunt, Chicken has two Nash equilibria. The Chicken region is separated into two regions depending on whether c_1 is positive or negative (Table 10).

Ells and Vello (1966) argue that the proper Chicken game (C1) should satisfy Eq. 21:

Table 10 Two variants of Chicken, one from each region

	0 (Give up)	1 (Continue)
0 (Give up)	$(5,5)$	$(2,6)$
1 (Continue)	$(6,2)$	$(1,1)$

Restricted Chicken (C1)

	0 (Give up)	1 (Continue)
0 (Give up)	$(3,3)$	$(2,6)$
1 (Continue)	$(6,2)$	$(1,1)$

Non-restricted Chicken (C2)

In C1 the players receive the highest payoff from strategy profile $\{0,0\}$, whereas in C2 more payoff is gained by alternating between $\{0,1\}$ and $\{1,0\}$ in iterated play

$$b + c < 2a \iff c_1 < 0. \tag{21}$$

The motivation behind this is the same as for the Prisoner's Dilemma. If $b + c > 2a$ the players will receive a higher payoff from alternating between action profiles $\{0,1\}$ and $\{1,0\}$ than by sticking to action 0. The idea behind Chicken is that cooperation is achieved when both players choose action 0, which is a different coordination problem than the C2 version. Rapoport et al. (1978) also show in their experiments that players relatively quickly find out how to take advantage of the fact that $b + c > 2a$. Hence there is support for distinguishing between C1 and C2.

5.4 Leader and Hero regions

The Leader and Hero games may be viewed as symmetric version of The Battle of the Sexes, where the two NE reside on the anti-diagonal rather than the diagonal (Harris, 1969; Rapoport et al., 1978; Rasmusen, 1994). None of the NE are Pareto Optimal, with the players preferring different NE. The difference between the games is that only in Hero is $|c_1| > |c_2|$, which makes the maximin strategy 0 in Leader and 1 in Hero (Table 11).

The names Leader and Hero stem from the slightly different coordination problems the games pose. In iterated versions of both games, a cooperative and fair strategy is for the players to alternate between $\{0,1\}$ and $\{1,0\}$. However, in want of such coordination, both players may choose their maximin strategies (0 in Leader; 1 in Hero). In this case, both players will receive a higher payoff if one of

Table 11 Leader and Hero games

	0 (Sacrifice)	1 (Exploit)
0 (Sacrifice)	$(1,1)$	$(3,5)$
1 (Exploit)	$(5,3)$	$(2,2)$

Hero

	0 (Exploit)	1 (Sacrifice)
0 (Exploit)	$(2,2)$	$(3,5)$
1 (Sacrifice)	$(5,3)$	$(1,1)$

Leader

In Leader, the first player to switch from the maximin strategy gets a benefit; in Hero the player who stays benefits

them switch action, but lower payoff if they both switch. Thus, in order for the players to increase their payoff, one of the players has to take initiative and change action, while the other waits. The difference between the games is that in Leader the player who shifts gets a higher payoff than the one who waits, but in Hero the switcher gets less payoff (Rapoport et al., 1978). Whence the names of the games.

Unsurprisingly perhaps, human players choose the maximin strategy much more frequently in Hero than in Leader (Rapoport et al., 1978). The Leader and Hero distinction is also included in Harris (1969) classification.

5.5 Conclusions

In this section we analyzed the standard game regions and compared to other relevant research. In the remaining regions we found that the tension between the common interest and self-interest is low, and it is therefore easy for the players to coordinate on the most favourable outcome. So we do not think any strategical insight is gained from analysing these games in further detail. Many of our regions have support from authors that conduct research on these games from starting points other than that of classifying. This therefore provides support for our classification.

6 Computer experiments

In this section we present an evolutionary computer experiment of the best strategies in various iterated symmetric 2×2 games. The design is somewhat inspired by the computer tournaments held by Axelrod and Hamilton (1981), though not restricted to Prisoner's Dilemma. The experiments provide some empirical evidence for our classification, showing that optimal strategies often change drastically between regions, while only undergoing continuous changes within regions.

6.1 Experiment design

The setup of the experiment is that strategies with memory length 1 compete on iterated symmetric 2×2 games in an evolutionary process. We describe each of the components in the following subsections.

6.1.1 Iterated games

By *iterated* 2×2 game we mean a 2×2 game that is played repeatedly for a given number of rounds. The players' payoffs will increase accumulatively in each iteration according to the same payoff function defined for the one-shot game.

6.1.2 Strategies with memory length 1

In Definition 6.1 below a formal definition of what we mean with *deterministic strategy with memory length one* is presented.

Definition 6.1 Let $G = (S, U, P)$ be a symmetric 2×2 game. A *strategy in iteration n* $s_{i,n} \in S_i$ is a strategy of player $i \in P$ in the n:th iteration of the game, $n \in \mathbb{N}$. We say that $\{s_{i,n}\}_{n\in\mathbb{N}}$ is *deterministic with memory length one* if $s_{i,0} = a \in A_i$ and $s_{i,n} = f_i(s_{i,n-1}, s_{-i,n-1})$ for $n > 0$, where $f_i : A_i \times A_{-i} \to A_i$.

Definition 6.1 states that a strategy is deterministic with memory length one if it has a deterministic starting action and if the action in the n:th iteration, $n > 0$, of an iterated game is uniquely determined by the outcome of the $n - 1$ iteration of the game. An example of a deterministic strategy of memory length one is presented in Example 6.2.

***Example* 6.2** In a symmetric 2×2 game an example of a deterministic strategy of length one is presented as in the Table 13. This strategy states that the player starts by playing action 0. If the outcome is $\{0, 0\}$ the action in the first iteration will be 0 again but if the outcome is $\{0, 1\}$ the action in the first iteration is 1. This strategy is in fact designed to imitate the last move of the opponent in every iteration. With 0 being the cooperative action and 1 the defective action, this strategy is referred to as *Tit for Tat* (Axelrod and Hamilton 1981).

Since $|A_i| = 2$ there are $|A_0 \times A_1| = 4$ possible outcomes in a symmetric 2×2 game. Player i has 2 possible actions to take as a response to each outcome and 2 possible starting positions and this means that there are $2^5 = 32$ possible deterministic strategies with memory length one. We denote the set of these strategies with S_D. Each of the two players can choose a strategy independently of the other player which means that there are $32^2 = 1024$ deterministic strategy profiles with memory length one for each symmetric 2×2 game.

6.1.3 Evolutionary process

The main steps of the evolutionary process is the following. Given an iterated game G and an initial population N of strategies with memory length one, the following steps repeats until stable statistics can be computed:

1. First, all strategies meet all other strategies in a round-robin tournament.
2. Based on the relative payoffs of the strategies, a *fitness* for each strategy is computed.
3. A new population N' is generated, where strategies in N are represented in proportion to their fitness, save for random mutations to the strategies.
4. Start over with Step 1 with the new population N'.

Population A population of strategies is simply a multi-set of deterministic strategies with memory length one. It is important to note that a population of strategies does not have to be a subset of the set of all deterministic strategies with memory length one, S_D, since we allow for the same strategy $s \in S_D$ to occur in a population more than once. Formally a *population of strategies* $\mathcal{N}$ is a multi-set of

strategies s such that $s \in \mathcal{N} \Rightarrow s \in S_D$. A population of order $n \geq 1$ is a population $\mathcal{N}$ with $|\mathcal{N}| = n$.

Fitness We need a measure that tells us how good a certain strategy performs in comparison to the other strategies in the population. This motivates Definition 6.3.

Definition 6.3 (*Total payoff*) Given a game G and a specific population of strategies $\mathcal{N}$, let p_{ij} denote the sum of payoffs over m games gained by strategy $s_i \in \mathcal{N}$ in a fight $F_m(s_i, s_j)$ against strategy $s_j \in \mathcal{N}$. We define the *total payoff of strategy i* as $p_i = \sum_{s_j \in \mathcal{N}} p_{ij}$.

We say that a strategy s_i is fitter than a strategy s_j in population $\mathcal{N}$ if $p_i > p_j$. However, just knowing if a strategy is fitter than another strategy is not enough for our purposes. Hence we introduce a measure of fitness in a population. We define the *fitness of strategy i with weight* α as

$$f_{\alpha_i} = \frac{|\mathcal{N}| \times e^{\alpha \times \text{score}_i}}{\sum_{s_j \in \mathcal{N}} e^{\alpha \times \text{score}_j}}.$$

Here the score of player i is the normalized payoff given by

$$\text{score}_i = \begin{cases} \dfrac{p_i - \min_{s_j \in \mathcal{N}} p_j}{\max_{s_k \in \mathcal{N}} p_k - \min_{s_j \in \mathcal{N}} p_j} & \text{if } \max_{s_k \in \mathcal{N}} p_k - \min_{s_j \in \mathcal{N}} p_j > 0 \\ 0 & \text{otherwise.} \end{cases}$$

The definition of fitness guarantees that for each strategy $s_i \in \mathcal{N}$, $0 \leq f_{\alpha_i} \leq |\mathcal{N}|$ and that $\sum_{s_i \in \mathcal{N}} f_{\alpha_i} = |\mathcal{N}|$, so the population size remains fixed. It also guarantees that if s_i is fitter than s_j, i.e. $p_i > p_j$ then $f_{\alpha_i} > f_{\alpha_j}$. The parameter α is used to alter the evolutionary pressure, with higher values of α the evolutionary pressure increases. We multiply with the size of the population, $|\mathcal{N}|$, so that the fitness of strategy i will correspond to the number of offspring strategy i gets in the following generation of strategies.

New population Given a game G, the idea is to start with the population of strategies $\mathcal{N} = S_D$ and let every pair of strategies fight each other, i.e. play G a given number of iterations. When every fight is done the fitness $f\alpha_i$ is calculated for every strategy $s_i \in \mathcal{N}$ and a new population $\mathcal{N}_{\text{new}}$ is generated such that the strategy $s_i \in \mathcal{N}$ occurs in the new population exactly $\lfloor f_{\alpha_i} \rfloor + 1$ times if the decimal part of f_{α_i} is greater than the decimal part of the fitness of the other strategies and $\lfloor f_{\alpha_i} \rfloor$ otherwise. This is to ensure that the $|\mathcal{N}_{\text{new}}| = |\mathcal{N}|$. We denote the number of *offspring* of strategy $s \in \mathcal{N}$ as $\sigma(s)$.

This means that strategies with high fitness relative to the other strategies get more offspring than the other strategies and weak strategies get less or no offspring. We also introduce the concept of *mutation*, denoted Υ, to the algorithm. When the new population is generated the mutation can randomly cause some of the strategies to mutate, i.e turn into some other strategy in S_D. This is done so that extinct strategies can always be reintroduced and challenge the dominating strategies.

This whole cycle is then repeated with $\mathcal{N} = \mathcal{N}_{\text{new}}$ as starting population a given number of generations $g \in \mathbb{N}$. The idea is that strategies that are good in the

particular game being played will dominate the majority of the populations given that g is large enough.

6.2 Execution

The algorithm has four hyper parameters, m, n, υ and α. In our particular case we set m, the number of rounds in a fight, to 50, n, the number of generations, to 1,000,000, υ, the probability of mutation, to 0.000001, and α, the evolutionary pressure, to 0.5. These parameters where set so that the convergence of the new populations where stable and gave room for mutations.

7 Results of computer experiments

In this section we describe and analyse the results of the experiments. Following a description of how we visualise the data in Sect. 7.1, we provide an of analysis action frequencies and strategy frequencies in Sect. 7.2. For a more detailed analysis, see Böörs and Wängberg (2017).

7.1 Visualisation of data

To be able to plot data from any 2×2 symmetric game in a finite 2d-plot, we visualise the data in a slightly different way than in Fig. 6. (In the Fig. 6 visualisation, games span all of $\mathbb{R}^2$.) Thus, rather than normalising games to $z_1 = 1$, we instead normalise games to the unit sphere

$$\{[c_1, c_2, z_1]^{\mathrm{t}} \in \mathbb{R}^3 : c_1^2 + c_2^2 + z_1^2 = 1\}.$$

We then apply *stereographic projection* (Coxeter and Greitzer 1967) to bring it down to $\mathbb{R}^2$, via

$$[c_1, c_2, z_1] \mapsto (c_1', c_2') = \left(\frac{c_1}{1 - z_1}, \frac{c_2}{1 - z_1}\right). \tag{22}$$

The projection is illustrated in Fig. 10.

As we show in Proposition 4.4 (Fig. 11), every game with $z_1 < 0$ is strategically equivalent to a game with $z_1 > 0$. Thus we can restrict our attention to games with $z_1 < 0$. These games all project inside the unit circle. Applying the equivalent

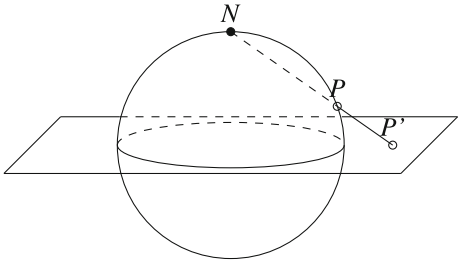


Fig. 10 The stereographic projection maps every game P on the three dimensional unit sphere onto the two dimensional plane

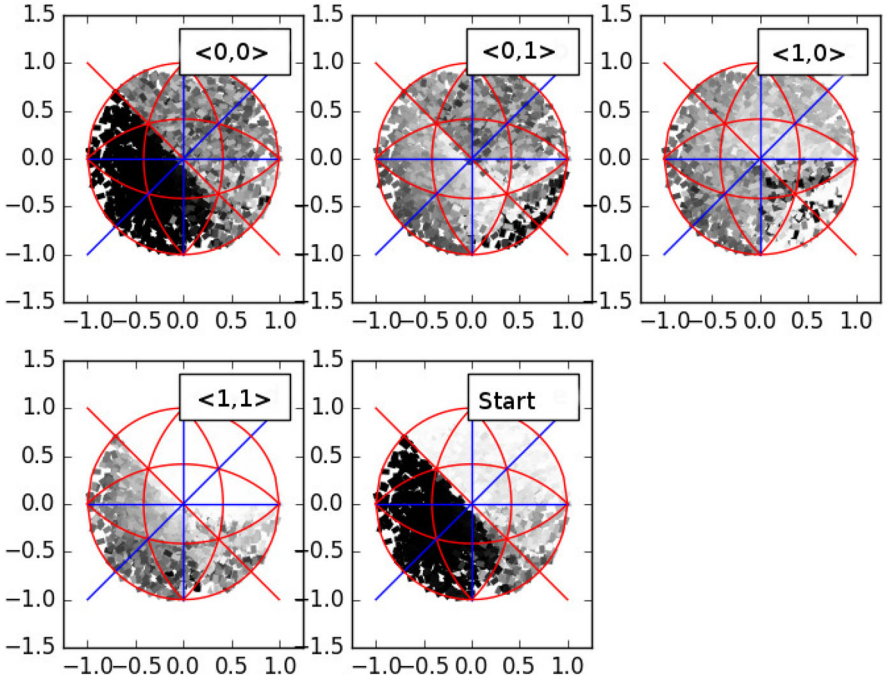


Fig. 11 Plots of action frequencies in the different contexts of Table 12. The colour encodes action frequency, with black for 100% action 0, and white for 100% action 1. The location encodes the payoff parameters, as described in Sect. 7.1

conditions as in Fig. 6 to the (c_1', c_2')-plane gives Fig. 12, where all regions fit inside the unit circle.

7.2 Analysis of results

By use of the generic algorithm presented in Sect. 6, we attempt to find the *evolutionary equilibrium* in the different game regions. This is to verify that the evolutionary equilibrium does not change abruptly inside regions, but only between regions, which would provide support for our classification. By evolutionary equilibrium we mean the limiting distribution of strategies from the described evolutionary process. When the limiting distribution does not exist, we are instead interested in the averaged distribution over the full cycle. We approximate the evolutionary equilibrium by taking the average distribution over the last 2000 rounds out of 1,000,000. Furthermore, we focus on 2 types of data in our analysis: (i) the frequency of actions in the five different contexts (past outcomes and start) displayed in Table 12 and (ii) the frequency of the 32 different strategies used in the experiments.

Visualisation of action frequencies is shown in Fig. 11. The colour encodes the action frequency, with black for 100% action 0, and white for 100% action 1; the location encodes the payoff parameters, as described in Sect. 7.1. The most

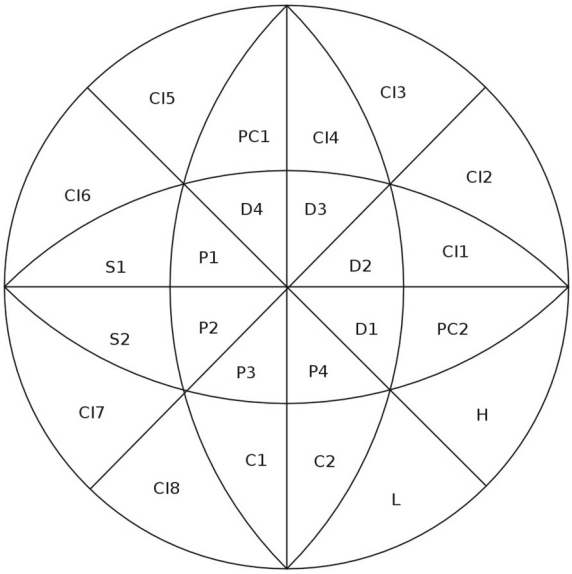


Fig. 12 In this map the game regions from Fig. 6 are shown after the stereographic projection

Table 12 A strategy with memory length one can be represented with a vector in S^5, with one action for each possible previous outcome, and one entry for the starting strategy

$\{0,0\}$	$\{0,1\}$	$\{1,0\}$	$\{1,1\}$	Start
s_{00}	s_{01}	s_{10}	s_{11}	s_{start}

important takeaway from the analysis of this data type is that there is no distinct change in action frequency inside any of our regions.

In Figs. 13 and 14 we provide the plots visualising the strategy frequencies. We only plot strategies that have a higher frequency than 15% to prevent the randomness of the mutation from interfering with our analysis. The result is that all classification conditions stated in Eqs. 9–13, except one, separate at least one type of strategy. We provide plots corresponding to Figs. 13 and 14 for all 32 strategies in Figs. 15 and 16 in Appendix 2 where we also observe that the evolutionary equilibrium does not change abruptly within any of our regions. The condition that fails to distinctively separate any strategies is the $c_1 \lesseqgtr c_2$ inequality. This separates regions where the c_1 and c_2 arrows point in the same direction which is a potential reason for why the relative strength between c_1 and c_2 does not make any important difference. In Table 14 we summarise which strategy of Figs. 13 and 14 is separated by what boundary.

The main conclusion we drew from the analysis of the computer experiments is that almost all conditions add interesting information about strategic play. Also

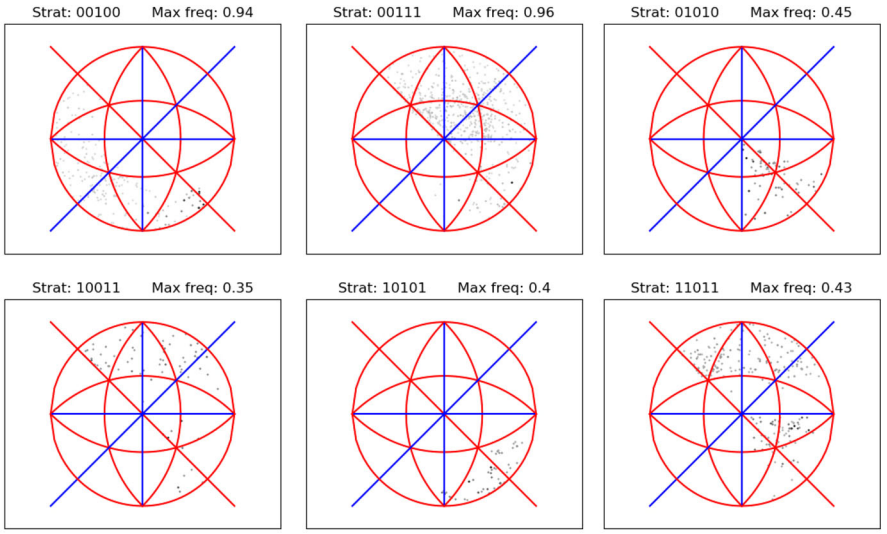


Fig. 13 In this figure the experimental result for strategies of memory length 1 (described in Tables 12 and 13) is plotted in gray scale in individual subplots. The label at the top of each subplot consists of the strategy represented in the plot as well as its maximum observed frequency in the experiment. The color of the dots (games) represents its frequency in the experiment. A dark color means that the strategy is frequent in the game. In every plot the frequencies have been normalized so that the game where the strategy is most frequent is black (colour figure online)

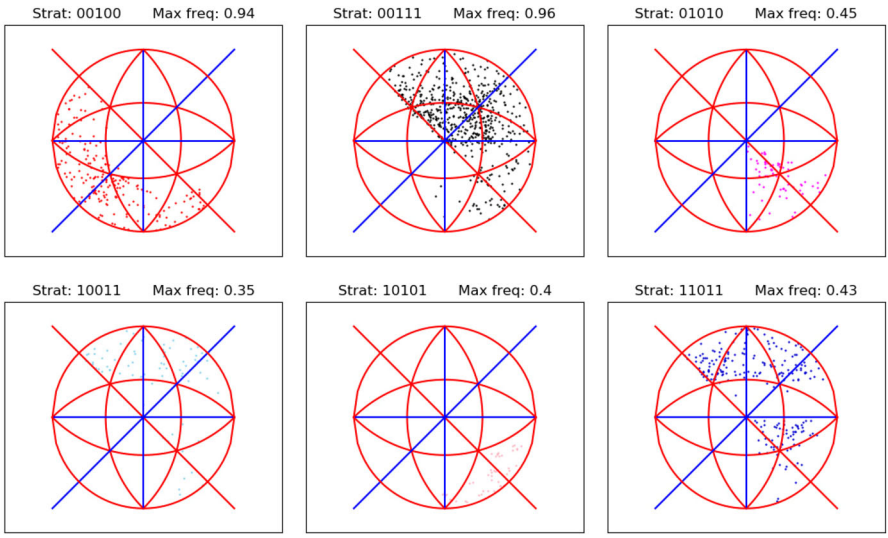


Fig. 14 In this figure the experimental result for strategies of memory length 1 (described in Tables 12 and 13) is plotted color in individual subplots. The label at the top of each subplot consists of the memory length one strategy represented in the plot along with its maximum frequency in the experiments (colour figure online)

Table 13 *Tit for tat* is an example of a deterministic strategy of length one

$\{0,0\}$	$\{0,1\}$	$\{1,0\}$	$\{1,1\}$	Start
0	1	0	1	0

Table 14 The table shows which classification boundaries each of the strategies shown in Figs. 13 and 14 is separated by

Boundary	Strategy
$c_1 = 0$	01010
$c_2 = 0$	11011
$c_1 = -c_2$	00111
$z_1 = c_1$	10101
$z_1 = -c_1$	00100
$z_1 = c_2$	10011
$z_1 = -c_2$	10101

there are no distinct borders in any of the figures presented earlier in this section where there is no line defined by some of our classification inequalities. This indicates that all important information is included in our classification. For a more detailed analysis, we refer to Böörs and Wängberg (2017).

8 Discussion

In this section we briefly summarize the paper and present a few open questions for further research.

8.1 Summary

A good theory for the classification of games is important for many reasons. It is a systematic way of studying games to understand what games are essentially different and why. It also provides an overview of possible strategical dilemmas, which can pose problems in many real-world situations (e.g. nuclear arms race). These dilemmas are thereby important to properly understand and be aware of.

In the beginning of the paper we reviewed existing classifications of games based on whether or not the classification conditions are well justified, in that it explains relevant observations, and have a simple and parsimonious structure. We considered the classification to be well justified if the authors provided some evidence to support their classification conditions. The motivation behind the reviewing desideratum of a simple structure is that simple theories generalize better and have a better track record. We found that none of the classifications reviewed have the

desired combination of well justified conditions and simple structure. This motivated us to define a classification of our own.

Building on the work of Kalai and Ehud (2013) we present the result that the payoff matrix of any 2×2 game can be decomposed into a non-conflict part and a zero-sum part. In all of the standard games the zero-sum games counteract the interests in the non-conflict games. This inspired us to create a partition of the space of symmetric 2×2 games based on the interest of the players in the decomposed parts of the games. We show that the resulting 24 classes capture every aspect captured by the reviewed authors that have a similar approach as us. We do, for example, provide results that state that every class has a specific type of NE and that every standard game has its own class. We also found support for our classes by comparing to other research. In addition, we prove that the concept of AE[5] used by Huertas-Rosero is captured by our classification, and that either all of the games in a class or none of them have a dominant strategy equilibrium. This supports the conjecture that the relevant strategic aspects of a game are determined by its decomposition.

To further justify the relevance of our classes we conducted computer experiments to see if the empirical data generated would support our class borders. Moreover, we create a compact 2-d map of our regions, unlike the previous classifications reviewed, enabling visualization of the results of our experiments and random sampling of games. The experiments showed that all but one of the conditions defining the classes made a difference for what strategies succeeded in the evolution-like setup in our experiments. This provides some support for our proposed classes. Perhaps even more important is the fact that the experiments did not suggest that we have missed any important distinction between games within any of our classes. Because the experiment only provides results concerning strategies with memory length one, we cannot make any conjectures about other types of strategies.

To conclude we add our classification to Table 4 which was used to summarize the literature review, see Table 15. Our classification is based only on the simplest relations between three parameters: c_1, c_2 and z_1. Hence we claim Desideratum 1. is satisfied. Furthermore, the parameters have an intuitive meaning, and are based on the conjecture that what makes a game interesting is the conflict working against the common interest. Furthermore, the regions are justified both by previous consensus on which games are interestingly different (Andreoni & Miller, 1993; Axelrod & Hamilton, 1981; Dubois et al., 2012; Harris, 1969; Lave, 1965; Radinsky, 1971; Rapoport & Albert, 1965; Rasmusen, 1994; Scodel, 1962; Skyrms, 2004), and computer experiments. Furthermore, the computer experiments and comparison to other research serves the purpose of justifying our regions. Desideratum 2. is thereby arguably satisfied.

[5] The Altruistic Equilibrium (AE) concept may not have any strategic relevance, but we mention this to stress that our classification captures all information that the other comparable classifications contain.

Table 15 In this table we add our classification to Table 4

Author	Desideratum 1: Simplicity and parsimony	Desideratum 2: Well-justified regions
Rapoport et al.	No	Yes
Robinson and Goforth (2003)	Yes	No
Harris (1969)	No	Yes
Borm (1987)	Yes	No
Huertas-Rosero (2003)	Yes	No
Böörs, Wängberg	Yes	Yes

Each classification receives a Yes if it manages to fulfill a desideratum and a No otherwise

8.2 Outlooks

In this work we have restricted to symmetric 2×2 games. Since many strategic interactions of interest involve more players and actions where the available actions of each player may differ, a natural continuation is to attempt to generalize this classification to all 2×2 games and beyond.

The decomposition that our classification is based on can also be generalized. For example, Candogan et al. (2011) show that games can be decomposed into a potential, a harmonic, and a non-strategic part of the game, where the harmonic and potential parts are related to the common interest and zero-sum components used in this paper.

In the case of symmetric 2×2 games, we can represent games with just three variables, c_1, c_2 and z_1 that are linear combinations of the payoffs. Our class boundaries define a relatively small number of classes and each class could be studied and verified in this article. Naturally, if one were to consider larger game spaces, the number of parameters required to represent a game would grow, and so would the number of classes. This can make it more difficult to study the properties of each class. For example, to represent symmetric 3×3 games using common interest variables and zero-sum variables analogous to the ones used to represent 2×2 games, one would need 5 common interest variables and 3 zero sum variables, that is 8 variables in total. This increase of dimensionality makes it more difficult to study the properties of each individual class. However, this can be an interesting topic for future work.

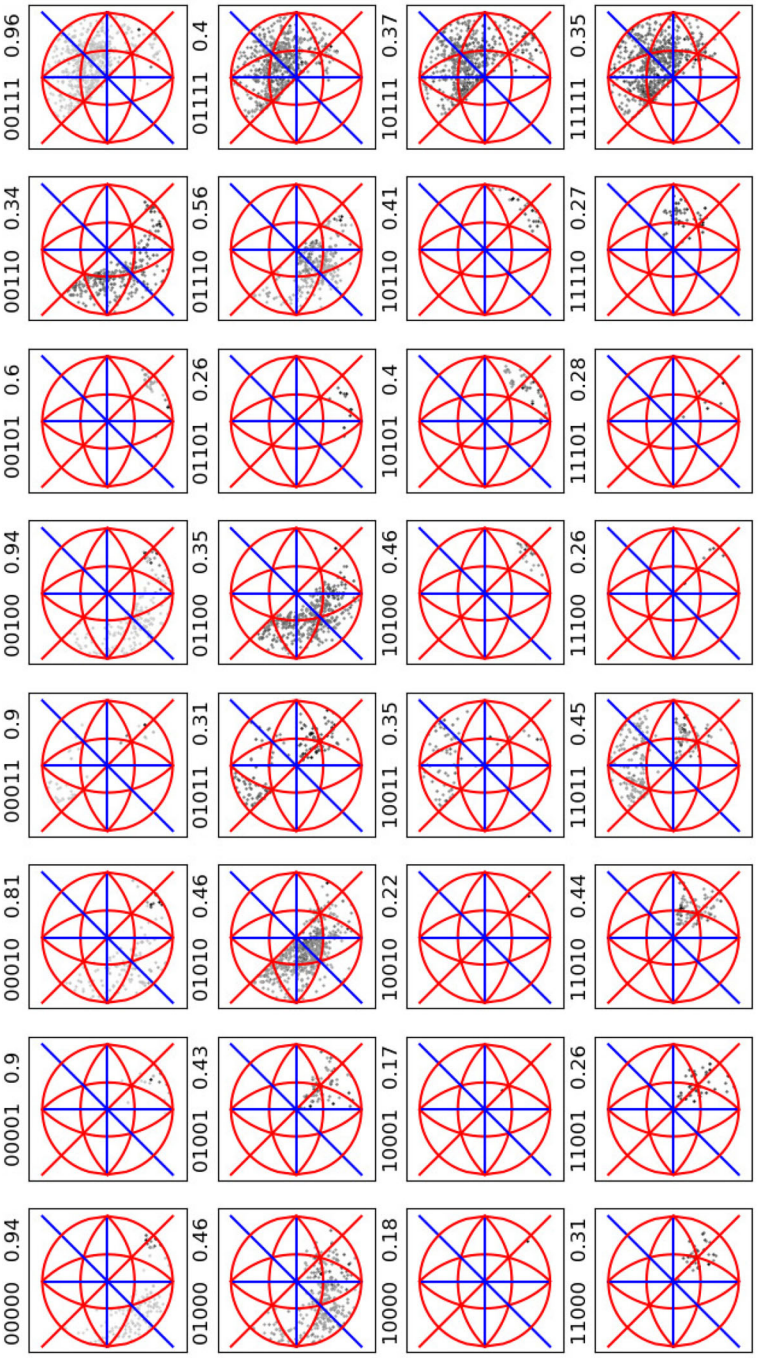
Appendix 1: List of notation

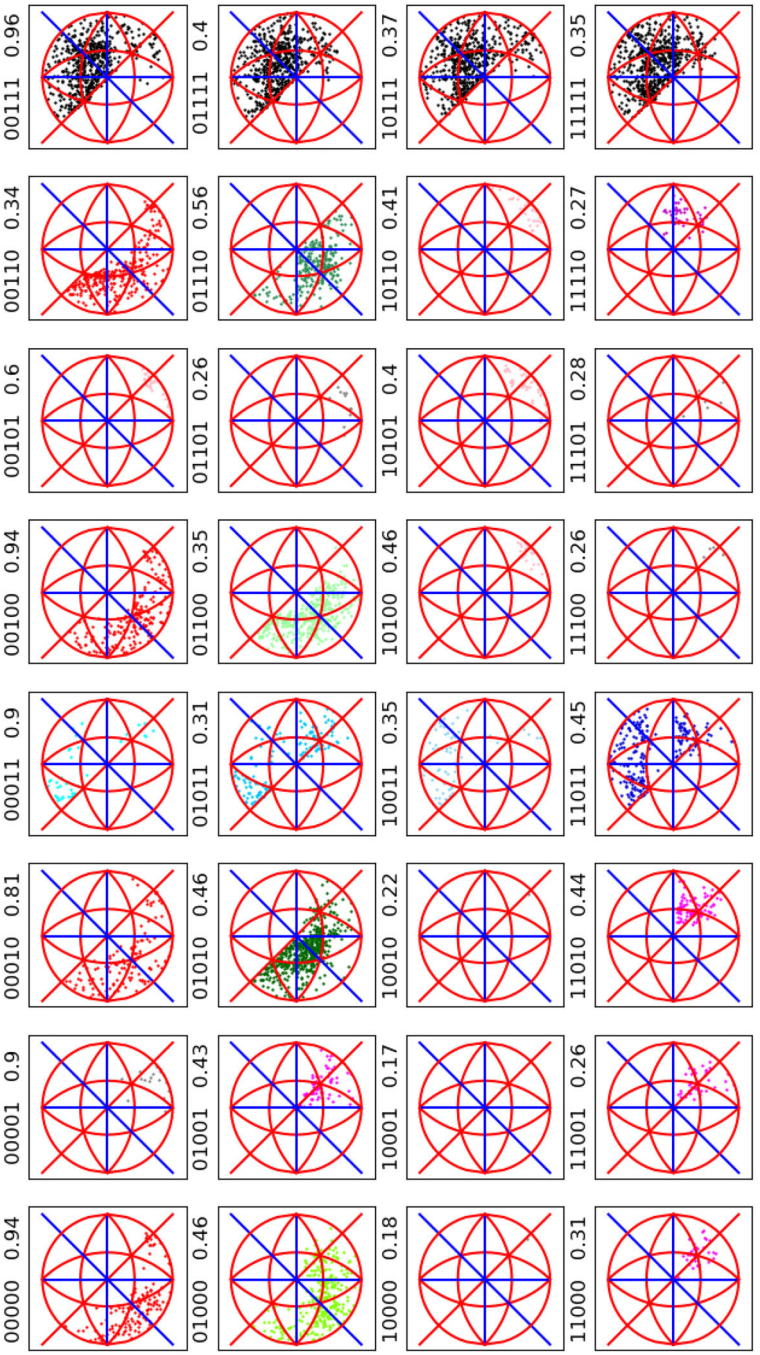
Symbol	Explanation
$\mathbb{R}, \mathbb{N}, \ldots$	Set of real,natural numbers
$k, l \in \mathbb{N}$	Indices for the natural numbers
J_n	The $n \times n$ matrix of ones
$i, j \in \mathbb{N}$	Indices for players
G	Game
$\mathcal{P}$	Set of players
u_i	Payoff function for player i
U	Set of payoff functions
A_i	Set of actions available to player i
a	Action profile
s_i	Strategy of player i
s_{-i}	Strategy profile for all players except player i
S_i	Set of strategies of player i
s	Strategy profile
S	Set of strategy profiles
P_i	Payoff matrix of player i
$\Gamma(G)$	The mixed extension of a game G
B_i	The best response/reply correspondence of player i
$\square$	End of proof
$\diamondsuit$	End of example
$\vec{v}$	Vector
$\sim$	Strategical equivalence
NE	Nash equilibrium
AE	Altruistic Equilibrium
z_k	Conflict parameter
c_k	Common interest parameter
a, b, c, d	Payoff parameters
$\{k, l\}$	Specific action/ strategy profile
v^{t}	Transpose of vector v

Appendix 2: Computer experiment results

See Figs. 15 and 16.

Fig. 15 In this figure the experimental result for strategies of memory length 1 (described in Tables 12 and 13) is plotted in gray scale in individual subplots. The label at the top of each subplot consists of the strategy represented in the plot as well as its maximum observed frequency in the experiment. The color of the dots (games) represents its frequency in the experiment. A dark color means that the strategy is frequent in the game. In every plot the frequencies have been normalized so that the game where the strategy is most frequent is black (colour figure online)


00000 0.94
01000 0.46
10000 0.18
11000 0.31
00001 0.9
01001 0.43
10001 0.17
11001 0.26
00010 0.81
01010 0.46
10010 0.22
11010 0.44
00011 0.9
01011 0.31
10011 0.35
11011 0.45
00100 0.94
01100 0.35
10100 0.46
11100 0.26
00101 0.6
01101 0.26
10101 0.4
11101 0.28
00110 0.34
01110 0.56
10110 0.41
11110 0.27
00111 0.96
01111 0.4
10111 0.37
11111 0.35


00000 0.94
01000 0.46
10000 0.18
11000 0.31
00001 0.9
01001 0.43
10001 0.17
11001 0.26
00010 0.81
01010 0.46
10010 0.22
11010 0.44
00011 0.9
01011 0.31
10011 0.35
11011 0.45
00100 0.94
01100 0.35
10100 0.46
11100 0.26
00101 0.6
01101 0.26
10101 0.4
11101 0.28
00110 0.34
01110 0.56
10110 0.41
11110 0.27
00111 0.96
01111 0.4
10111 0.37
11111 0.35

◀**Fig. 16** In this figure the experimental result for strategies of memory length 1 (described in Tables 12 and 13) is plotted color in individual subplots. The colours have been chosen so that strategies that are strong in similar areas have the same colour. The label at the top of each subplot consists of the memory length one strategy represented in the plot along with its maximum frequency in the experiments (colour figure online)

Appemdix 3: Dominant strategy equilibria

In this section we prove that the any game residing in the interior of one our classes have a dominant strategy equilibrium if and only if the games in the class have a unique NE. This is not true for games on the border between two classes. To make the presentation more clear, we use two lemmas to prove the main theorem, Theorem 11.3.

Lemma 11.1 (Unique Nash in symmetric games) *Let G be a symmetric 2×2 game with distinct payoffs, i.e. no pair of payoffs are equal. Then G has a dominant strategy equilibrium if and only if G has a unique NE. Also, the NE is the dominant strategy equilibrium.*

Proof Suppose that the payoff matrix of a game is the following.

		Player 2	
		0	1
Player 1	0	(a, a)	(b, c)
	1	(c, b)	(d, d)

Suppose further that a, b, c and d are distinct so that no pair of payoffs are equal. Then Lemma 4.6 states that

1. $\{0, 0\}$ is NE iff $a > c$.
2. $\{0, 1\}$ is NE iff $b > d$ and $c > a$.
3. $\{1, 0\}$ is NE iff $c > a$ and $b > d$.
4. $\{1, 1\}$ is NE iff $d > b$.

We will use the best response correspondence (Definition 2.10), to prove the lemma. Note that a strategy $s_1 \in S_1$ is a strictly dominant strategy for player 1 if and only if $B_1(0) = B_1(1) = \{s_1\}$. Also, in a symmetric game there is no difference between the best response correspondences of the two players, i.e. $B_1 = B_2$ and hence it is enough to consider player 1 in this proof.

Of course, 2 is true if and only if 3 is true. So neither $\{0, 1\}$ nor $\{1, 0\}$ can be a unique NE. That leaves us with two possible unique NE.

Suppose $\{0, 0\}$ is a unique NE so that 1 is true but 2,3 and 4 are false. Then we know that $a > c$ and $d < b$ and hence if player 2 plays 0, player 1 will play 0 and if player 2 plays 1 player 1 will play 0. This can be expressed as $B_1(0) = B_1(1) = \{0\}$. That is playing 0 is a strictly dominant strategy for player

1. Because of symmetry, the exact same reasoning is true for player 2. In this case 0 is a strictly dominant strategy for both players. This proves that if $\{0,0\}$ is a unique NE, then $\{0,0\} \in S$ is a strictly dominant strategy profile. According to Definition 2.9, this means that $\{0,0\}$ is a dominant strategy equilibrium.

Now suppose that $\{1,1\}$ is the unique NE so that 1, 2 and 3 are false but 4 is true. This means that $a < c$ and $d > b$ and therefore $B_1(0) = B_1(1) = \{1\}$. From this we can tell that $\{1,1\} \in S$ is a strictly dominant strategy profile and that $\{1,1\}$ is a dominant strategy equilibrium.

There are no other possible unique NE outcomes. Therefore if a 2×2 symmetric game with distinct payoffs has a unique NE, then that outcome is also a dominant strategy equilibrium.

Suppose now that the game has two NE. This means that either $\{0,1\}$ and $\{1,0\}$ or $\{0,0\}$ and $\{1,1\}$ are the two NE. If $\{0,1\}$ and $\{1,0\}$ are the NE, then we know that $b > d$ and $c > a$. Hence $B_1(0) = \{1\}$ and $B_1(1) = \{0\}$. Since $B_1(0) \neq B_1(1)$ there is no strictly dominant strategy profile in the game and as a result there is no dominant strategy equilibrium. Similarly, if $\{0,0\}$ and $\{1,1\}$ are the NE, then $a > c$ and $d > b$. Hence $B_1(0) = \{0\} \neq B_1(1) = \{1\}$ and there is no strictly dominant strategy profile.

We have proved that if there is a unique NE in a symmetric 2×2 game with distinct payoffs, then the NE outcome is also a dominant strategy equilibrium. Also, if such a game has two NE, then there is no dominant strategy equilibrium. Since a symmetric 2×2 game with distinct payoffs has either one or two NE, we have that such games have a dominant strategy equilibrium in an outcome x iff x is a unique NE. □

Next we need a result that states that every game that belong to exactly one of our 24 classes have distinct payoffs. Lemma 11.2 provides such result.

Lemma 11.2 (Distinct payoffs) *Let G be a symmetric 2×2 game. If G belongs to exactly one class in our classification, then the payoffs of G are distinct.*

Proof In Eq. (17) we show that the following holds.

$$\begin{aligned} a &= \frac{1}{4}(2x - 3c_1 - c_2) \quad b = \frac{1}{4}(2x + c_1 - c_2 + 4z_1) \\ c &= \frac{1}{4}(2x + c_1 - c_2 - 4z_1) \quad d = \frac{1}{4}(2x + c_1 + 3c_2). \end{aligned} \tag{23}$$

This means that $a = b \iff z_1 = -c_1$ so any game with $a = b$ will lie on the border $z_1 = -c_1$ between two or more classes (see Fig. 6 on page 15). In the same way we have:

$$\begin{aligned} a = c \iff z_1 = c_1 \quad a = d \iff c_1 = -c_2 \quad b = c \iff z_1 = 0 \\ b = d \iff z_1 = -c_2 \quad c = d \iff z_1 = c_2. \end{aligned}$$

That is, whenever two payoffs are equal, the game lies on the border between two of our classes. □

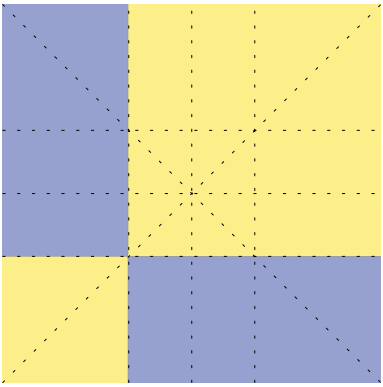


Fig. 17 The yellow regions contain games that have a dominant strategy equilibrium and the blue regions contain games that does not (colour figure online)

Now we present an important theorem that divides our 24 classes into two categories based on if the classes contain games that have a dominant strategy equilibrium or not.

Theorem 11.3 (Existence of dominant strategy equilibria) *Let G be a symmetric 2×2 game such that G belong to exactly one class in our classification. Then G has a dominant strategy equilibrium if and only if G has a unique NE. Also the NE is the dominant strategy equilibrium.*

Proof Let G be a game such as described in Theorem 11.3. Lemma 11.2 states that G has distinct payoffs and Lemma 11.1 states that G has a dominant strategy equilibrium in outcome x if and only if G has a unique NE in x. □

In Fig. 17 classes of games that have a dominant strategy equilibrium are coloured yellow and classes that does not have a dominant strategy equilibrium are coloured blue. Remember that games on the border between two or more classes are not included in Theorem 11.3.

Acknowledgements We thank the reviewers for many helpful comments and suggestions for improvement.

Funding Open access funding provided by Stockholm University.

References

Andreoni, J., & Miller, J. H. (1993). Rational cooperation in the finitely repeated prisoners dilemma: Experimental evidence. *The Economic Journal, 10327*(418), 570–585. ISSN:00220531. https://doi.org/10.1016/0022-0531(82)90029-1. arXiv:1011.1669v3

Axelrod, R., & Hamilton, W. D. (1981). *The evolution of cooperation* (Vol. 211, No. 4489, pp. 1390–1396). Basic Books. ISBN:0036807518110. https://doi.org/10.1086/383541. arXiv:t8jd4qr3m [13960]

Baker, A. (2007). Occams Razor in science: A case study from biogeography. *Biology & Philosophy, 22*(2), 193–215.

Böörs, M., & Wängberg, T. (2017). Classification by Decomposition: A partitioning of the space of 2×2 symmetric games. http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-137991. Accessed 15 Nov 2021

Borm, P. (1987). *A classification of* 2×2*bimatrix games*. Katholieke Universiteit Nijmegen, Mathematisch Instituut. https://pure.uvt.nl/portal/files/633023/PB09.PDF

Candogan, O., Menache, I., Ozdaglar, A., & Parrilo, P. A. (2011). Flows and decompositions of games: Harmonic and potential games. *Mathematics of Operations Research, 36*(3), 474–503.

Coxeter, H. S. M., & Greitzer, S. L. (1967). Geometry revisited. In: *Media* (p. 207). https://doi.org/10.1007/978-0-387-79148-7. http://www.librarything.com/work/175563/book/28044691

Dubois, D., Willinger, M., van Nguyen, P. (2012). Optimization incentive and relative riskiness in experimental stag-hunt games. *International Journal of Game Theory, 41*(2), 369–380. ISSN:00207276. https://doi.org/10.1007/s00182-011-0290-x. arXiv:cs/9605103. http://www.lameta.univ-montp1.fr/marcwillinger/textes/coordination.pdf

Ells, J. G., & Sermat, V. (1966). Cooperation and the variation of payoff in non-zero-sum games. *Psychonomic Science, 5*(4), 149–150. ISSN:0033-3131. https://doi.org/10.3758/BF03328325

Gonzalez-Diaz, J., García-Jurado, I., & Fiestras-Janeiro, M. G. (2010). *An introductory course on mathematical game theory* (Vol. 115, p. 324). ISBN:9780821851517.

Harris, R. J. (1969). A geometric classification system for 2×2 interval-symmetric games. *Behavioral Science*. http://search.proquest.com/openview/baa208f8d0041f0464a7c6da8cd66d13/1?pq-origsite=gscholar%7B%5C&%7Dcbl=1818492.

Huertas-Rosero, Á. F. (2003). A cartography for 2×2 symmetric games. In: *III Colombian Congress and I Andean international conference of operational research* (p. 13). arXiv:cs/0312005

Huertas-Rosero, Á. F. (2004). Classification of quantum symmetric nonzero-sum 2×2 games in the Eisert scheme, p. 154. arXiv:quant-ph/0402117v2

Kalai, A., & Kalai, E. (2013). Cooperation in strategic games revisited. *The Quarterly Journal of Economics, 128*(2), 917–966.

Lave, L. B. (1965). Factors affecting cooperation in the prisoners dilemma. *Behavioral Science, 10*(1), 26–38. ISSN:10991743. https://doi.org/10.1002/bs.3830100104

Radinsky, T. L. (1971). Exposing an individual to two types of prisoners dilemma game matrix formats. *Psychonomic Science, 24*(2), 62–64. ISSN:00333131. https://doi.org/10.3758/BF03337894

Rapoport, A., & Chammah, A. M. (1965). *Prisoners dilemma: A study in conflict and cooperation* (p. 258). University of Michigan Press. ISBN:0472061658.

Rapoport, A., Guyer, M. J., & Gordon, D. G. (1978). The 2×2 game. *Philosophy and Phenomenological Research, 39*(2), 292. ISSN:00318205. https://doi.org/10.2307/2106992

Rasmusen, E. (1994). *Games and information* (2nd ed.). Blackwell.

Robinson, D., & Goforth, D. (2003). A topologically-based classification of the 2×2 games. In: *37th Annual CEA conference 30.1966*. http://economics.ca/2003/papers/0439.pdf

Scodel, A. (1962). Induced collaboration in some non-zero-sum games induced collaboration in some nonzero-sum games1. *Source: The Journal of Conflict Resolution, 6*(4), 335–340. ISSN:0022-0027. https://doi.org/10.1177/002200276200600404. http://www.jstor.org/stable/172610%7B%5C%%7D0Ahttp://www.jstor.org/page/info/about/policies/terms.jsp%7B%5C%%7D0Ahttp://www.jstor.org

Skyrms, B. (2004). *The stag hunt and the evolution of social structure*. Cambridge University Press.