Understanding Human Behavior via Similarity: A Geometric and Behavioral Rules-Based Approach to Games*

Amil Camilo Moore† Fabrizio Germano‡ Rosemarie Nagel§

December 23, 2025

Abstract

We study similarity in the complete set of one-shot two-player 2×2 games with payoffs from $\{1, 2, 3, 4\}$ without replacement. Similarity is defined geometrically via a neighborhood structure on games and continuity of behavior, and is applied to both theoretical rules, such as Nash equilibrium or level-k reasoning, and to experimental data. This yields theoretical and empirical similarity classes of games. We conduct a large-scale experiment, in which each subject plays all 144 games in our class. We find that empirically inferred similarity classes diverge sharply from those predicted by Nash equilibrium and dominance reasoning. Yet, they align closely with a level-k variant, with deviations reflecting fairness and efficiency concerns. At the individual level, subjects' play can be classified according to primary and secondary rules, conforming with either level-k ($0 \leq k \leq 5$) or a fairness and efficiency-based heuristic. The main insights extend to strategic settings beyond our 2×2 games.

*We thank Larbi Alaoui, Jose Apesteguia, Elia Benveniste, Ayah Bohsali, Luke Boosey, Victoria Cirlot, Vince Crawford, John Duffy, Pia Ennuschat, Ido Erev, Vinicius Ferraz, Evan Friedman, Alexander Frug, Malachy Gavan, Ben Golub, Duarte Gonçalves, Glenn Harrison, R. Mark Isaac, Annie Liang, Raquel Lorenzo, Antonio Penta, Yaroslav Rosokha, Al Roth, Andrea Salvanti, Patrick Sewell, Colin Sullivan, Elias Tsakas, Robert Wojciechowski, and audiences in seminars and confrences in Barcelona, Bielefeld, Budapest, Lancaster, London, New Orleans, Stony Brook and Tucson for insightful comments. Germano and Nagel acknowledge financial support from Grants respectively MICIU/AEI/UE-PID2023-153318NB-I00 and MICIU/PID2021-125538NB-I00 and from the Severo Ochoa Programme for Centres of Excellence in R&D (Barcelona School of Economics CEX2024-001476-S), funded by MCIN/AEI/10.13039/501100011033. All mistakes are our own.

†Department of Economics and Business, Universitat Pompeu Fabra, amilcar.camilo@upf.edu

‡Department of Economics and Business, Universitat Pompeu Fabra, and BSE, fabrizio.germano@upf.edu

§Department of Economics and Business, Universitat Pompeu Fabra, and BSE, rosemarie.nagel@upf.edu

1 Introduction

Similarity is a fundamental organizing principle of human cognition. People perceive, classify, and respond to the world not by analyzing every situation anew but by relating new experiences to familiar ones that appear similar. In perception and categorization, two objects are typically regarded as similar if they share the same features (Tversky, 1977) or if they are close in space (Shepard, 1987). These relations guide how humans organize information, form expectations, and make choices across different environments (Roads and Love, 2024). Yet, despite its centrality to cognition in individual decision-making, a systematic study of behavioral similarity across a wide variety of games has not been done.[1] Game theory provides types of games based on logical or equilibrium properties. Still, it remains unclear how games are to be classified based on actual behavior.[2]

We systematically study similarity between games both theoretically and experimentally, based on the games' payopff structure and behavioral rules, as well as actual behavior in the most basic strategic environment: the full set of one-shot 2×2 games, in which two players choose between two actions and payoffs are drawn without replacement from the set $\{1, 2, 3, 4\}$.[3] Despite its simplicity, this environment serves as a microcosm of strategic interaction, capturing fundamental structures underlying economic behavior. It encompasses cooperation, defection, coordination, punishment, and purely competitive interactions, and generalizes naturally to canonical environments such as ultimatum, n-person coordination, public goods, Cournot, and beauty-contest games. By studying the entire class of such games with a simple payoff structure, we systematically analyze how small changes in payoffs predict continuous or discontinuous shifts in behavior, and relate it to game-theoretic principles and behavioral rules. This approach provides an atlas of gameplay and similarity classes for the most basic games, revealing, as a byproduct, understudied games—such as neighbors of the prisoner's dilemma or even mini-ultimatum games. Critically, our results establish clear principles for when Nash equilibrium is or is not a robust predictor of behavior—for instance, in coordination games with Pareto-ranked equilibria—offering foundational insights applicable to more general strategic settings. [4]

[1] Rubinstein (1988), Aizpurua et al. (1990), and Gilboa and Schmeidler (1995) study the similarity of judgments in individual decision-making theoretically. Evers et al. (2022) study similarity through categorization in mental accounting. Jehiel (2004) introduces analogy-based expectation equilibrium for multi-stage games with perfect information. Agents bundle other agents' decision nodes into analogy classes and learn only the average behavior within each class. To our knowledge, all experimental similarity studies on games have chosen a small subset of game classes; see, e.g., Guida, and Devetag, 2013 on prisoner's dilemma and stag hunt games, or Grimm and Mengel (2012) on acyclic 3×3 games with one equilibrium and the related literature therein.

[2] Examples of properties are the number of equilibria, the distinction between unique, multiple, or mixed equilibria, dominance solvability (Gale, 1953; Raiffa and Luce, 1957), games of strategic substitutes and complements (Vives, 2005), the distinction between zero-sum and non-zero-sum (von Neumann, 1928), or cooperative and non-cooperative games (von Neumann and Morgenstern, 1944, Nash, 1950, 1952).

[3] Rapoport, Guyer, and Gordon (1976) were the first to classify 2×2 games according to pure conflict (zero-sum games or games with no pure equilibrium), mixed motives (like prisoner's dilemma game), and common interest games. Robinson and Goforth (2005), Bruns (2015), and Shubik (2012) analyze the same 2×2 games we study, primarily classifying games in terms of traditional game theory concepts as preference ordering, dominance structure, number of NE or Pareto efficient outcomes. We draw especially on the work by Rapoport, Guyer, and Gordon (1978) and Shubik (2012).

[4] Recent work applying machine learning algorithms to games demonstrates how the structural analysis of simple games can be generalized far beyond small classes and how 2×2 constitute important foundational blocks, see e.g., Omidshafiei et al. (2020)and Biggar and Shames (2023).

The experimental literature started by selecting specific games prone to deviations of human behavior from Nash equilibrium, uncovering parsimonious behavioral rules.[5,6] While these approaches have significantly advanced our understanding of behavior, they raise important questions about generalizability. Specifically, how well do insights derived from selected experimental games extend to a broader domain of games? Addressing this question involves several challenges. One approach is to identify patterns across existing experimental studies, as seen in meta-analyses, for instance, by Wright and Leyton-Brown (2014), who examine behavior in 3×3 games.[7]

Another approach consists in generating a large diversity of games within a given class, with randomly drawn payoffs, typically from an interval such as [0, 100], and with the same aim as before, to uncover behavioral rules: 2×2 games (Zhu et al., 2025), 3×3 games (Selten et al., 2002, Fudenberg and Liang, 2019, FL hereafter), or constant-sum 3×3 games (Erev and Roth, 1995). The major drawback of randomly generating payoffs is that it omits strategically significant games that might be statistically rare, e.g., symmetric games, or games with cells containing equal payoff for both players (i.e., games of mixed or common interest), thereby eliminating essential games.[8] Indeed, these studies of random games predominently reveal level-k patterns (with k=1 most prominent) and, at times, a preference for maximizing joint payoffs, yet, rarely social preference types. FL observed such restricted behavioral outcomes and therefore generated symmetric games that did not exhibit level-k as modal behavior prediction through machine learning tools. This principle led to games allowing for Pareto-optimal equal-payoff cells and to the discovery of a new rule, the Pareto-dominant NE. However, Camilo and Nagel (2026) show that the better description of this aggregate behavior is a (near)-equal split rule, as developed in this present paper. None of these papers studies a complete domain of a game class that is capable of examining the relationship between game structure and behavioral rules as we will show in this paper.

Ert et al. (2011) is closely related to this paper, as it studies the simplest two-player extensive-form game with binary choices for both players with carefully constructed games drawing payoffs from a small reduced payoff-set $\{-8, -7, ..., 0, ..., 8\}$. The resulting games are structured according to whether a game is trivial (with aligned preferences between both players) or having actions as (costly) punishment, dictator outcomes, or other psychological interpretations of payoff relationships. They do not provide a classification based on subjects' aggregate behavior, as we do, nor a classification based on Nash or level-1 concepts.

[5] Some examples are ultimatum games (Gueth et al. 1982), beauty contest games (Nagel, 1995), 2×2 games with unique mixed equilibrium games (Erev and Roth, 1998), 3×3 games (Stahl and Wilson, 1994, 1995, Costa-Gomes, Crawford, and Broseta., 2001), and a diverse set of games by Goree and Holt, 2001.

[6] Heterogeneity of human behavior in games has motivated numerous solution concepts in behavioral game theory: social preferences (Bolton and Ockenfels, 2000; Fehr and Schmidt, 1999), team reasoning (Sugden, 1993; Bacharach, 2006), level-k and cognitive-hierarchy models (Nagel, 1995; Stahl and Wilson, 1995; Camerer et al., 2004; Fudenberg and Liang, 2019), and quantal-response equilibrium (McKelvey and Palfrey, 1995), see also the survey by Crawford et al. (2013); and learning models by Roth and Erev (1995), and Camerer and Ho (1998), are not discussed here because of the "initial play" nature of our experiment.

[7] Güth and Kocher (2014) review over three decades of variations of ultimatum bargaining experiments, highlighting behavioral motives and design variations. Billinger and Rosenbaum (2023) provide a meta-analysis of design variables that influence cooperation in public goods games. Mauersberger and Nagel (2018) discuss levels of reasoning in generalized Keynesian beauty-contest games.

[8] Alon et al. (2025) further document biases and regularities, in particular in connection with dominance solvable games, when working with randomly drawn games.

To formalize similarity in the context of our 2×2 games, we adopt Germano's (2006) geometric framework, which organizes games based on their neighborhood structure and the continuity of choices prescribed across games. We extend this approach by applying it, for the first time, to behavioral rules and experimental data.[9] Specifically, for any rule or list of choices defined on a given space of games (e.g., Nash equilibrium actions over 2×2 games), two games are similar if there exists a path of neighboring games such that the choices made at the games remain continuous along the path. The discontinuities of the choices partition the space of games into a (finite) number of connected components of games, which are further reduced in number after allowing for operations of relabeling actions and/or players, and which we finally refer to as *similarity classes*. The resulting representation of the game space with its neighborhood structure aligns with Shepard's spatial similarity model, yielding a single space of (neighboring) games. Within this space, however, distinct heuristics or behavioral rules induce different meaningful classes. For instance, level-k reasoning emphasizes strategies that maximize expected payoffs under assumptions of limited opponent rationality, while fairness-based reasoning prioritizes outcomes that (nearly) equalize payoffs across players. These heuristics rely only on a subset of the information in a payoff matrix. Such selective attention echoes Tversky's (1977) feature-based model of similarity, in which objects are grouped based on salient attributes rather than their complete descriptions.

Our analysis identifies key distinctions between the similarity classes derived from the various theoretical rules. The level-k variant produces four distinct similarity classes based on how different levels of reasoning converge to equilibrium. In a first class, both players' level-1 actions already constitute a Nash equilibrium (this class contains all two-sided dominance games, some one-sided dominance games and some coordination games); in a second class one player's level-1 action and the other player's level-2 action constitute the unique Nash equilibrium (this class contains the remaining one-sided dominance games); and two more classes, where either the two players' respective level-1 and level-2 actions constitute two different Nash equilibria (this class contains the remaining coordination games) or where no combination of level-k actions constitutes a Nash equilibrium (this contains all matching pennies type games, that have no pure Nash equilibria). This classification differs significantly from the Nash equilibrium one. Under Nash equilibrium, double-dominance and one-sided dominance games are classified together, while coordination games and matching pennies type games are classified into two separate groups. Similarity classes based on (strict) dominance separate games with two-sided dominance from those with one-sided dominance, and the remaining games (coordination and matching pennies type) are classified as with Nash equilibrium. These differences can be tested empirically.

To test the empirical relevance of the different rules and corresponding theoretical similarity classes, we conduct a large-scale experiment in which each human subject plays all the mentioned 2×2

[9] While Germano (2006) studied spaces of finite games as Euclidean spaces with corresponding notions of continuity of correspondences (see Aliprantis and Border, 2006), we focus on the space of 2×2 games with payoffs 1, 2, 3, 4. In our case, two games are neighbors if they are close in the discrete space, meaning that their Euclidean distance is at most $\sqrt{2}$. This is equivalent to requiring that one game can be obtained from the other by swapping the payoffs of one player, when the difference between the payoffs swapped is exactly 1; see Definition 1. Continuity of rules also simplifies to the constancy of actions or sets of actions recommended by the rules.

games once, without receiving feedback from the plays of other subjects. Thus, it follows the literature of initial play (one-shot). Additionally, having each subject play all 78 games in both player roles (a total of 144 choice-situations we refer to as perspectives) — represents the largest set of distinct game choices in the experimental literature. This enables a detailed analysis of individual-level behavior, revealing systematic differences in how subjects perceive and respond to different games. Notably, we find that many subjects employ secondary rules, which are often labeled as errors in traditional experiments with 10–20 game choices. This exhaustive experimental design not only addresses the challenges of generalizability but also deepens our understanding of both aggregate and individual decision-making across the full domain.

Within this large-scale experimental framework, we examine the full space of such games, grouping them according to both aggregate behavioral data across all subjects or subgroups and theoretical principles, including concepts such as Nash equilibrium, strict dominance, and behavioral models such as level-k reasoning or fairness considerations. We then compare the empirically derived classifications with those implied by theory, thereby identifying where observed patterns of play converge with or deviate from established models of strategic reasoning. By systematically studying all possible payoff configurations in this minimal strategic setting, we obtain a comprehensive map of the 2$\times$2 interaction landscape. This allows us to characterize environments ranging from highly competitive to strongly cooperative, and to relate these structural features to individual differences in strategic sophistication, belief hierarchies, and fairness orientations. Our overarching objective is to demonstrate that organizing games according to their empirical similarity classes provides a robust and predictive framework for understanding behavior across the full spectrum of strategic environments. Predictions in a game—or within a class of games—are straightforward when different theoretical and behavioral models prescribe the same action for both player roles. This occurs in 31% of games, which we refer to as one-outcome games, in which the Nash equilibrium (NE), fairness, and level-k rules coincide. Similarly, in some games, only one of the two player roles has a uniquely prescribed action across these rules. In all other cases, at least one player role is subject to conflicting prescriptions across the models.

Empirical similarity classes, derived from the aggregated behavior of subjects, most closely align with the level-k-based classification. This is unsurprising, as the majority of subjects' behavior follows level-k reasoning. However, deviations arise in games in which level-k and Nash equilibrium strategies prescribe outcomes that are neither fair nor efficient, despite the existence of such outcomes. In these cases, some subjects who predominantly follow level-k reasoning adopt a secondary rule, which we term the (near) equal-split rule. This rule prioritizes fairness and efficiency and is also observed among subjects who primarily make fairness-driven decisions. Notably, fairness-oriented subjects often use level-k reasoning as a secondary approach.

These findings reveal important structural distinctions between level-k reasoning, Nash equilibrium, and considerations of efficiency and fairness. They emphasize how differences in reasoning steps, dominance patterns, and fairness considerations influence both theoretical and empirical similarity

$\mathbf{g_{58}}$	A	B	
A	$\underline{4}, \underline{4}$	$2, 3$	**0.82**
B	$1, 1$	$\underline{3}, \underline{2}$	**0.18**
	0.50	**0.50**	

$\mathbf{g_{59}}$	A	B	
A	$\underline{4}, \underline{4}$	$2, 2$	**0.87**
B	$1, 1$	$\underline{3}, \underline{3}$	**0.13**
	0.50	**0.50**	

$\mathbf{g_{60}}$	A	B	
A	$\underline{4}, \underline{4}$	$2, 1$	**0.85**
B	$1, 2$	$\underline{3}, \underline{3}$	**0.15**
	0.85	**0.15**	

$\mathbf{g_{70}}$	A	B	
A	$\underline{4}, \underline{4}$	$1, 2$	**0.52**
B	$3, 1$	$\underline{2}, \underline{3}$	**0.48**
	0.52	**0.48**	

Figure 1: The four coordination games g_{58}, g_{59}, g_{60}, and g_{70} with empirical frequencies of play. Underlined payoffs indicate Nash equilibrium outcomes.

across game types. While the level-k model provides the best match for empirical similarity classes, incorporating efficiency and fairness considerations is necessary to account for certain deviations in behavior.

To illustrate our approach, consider the four coordination games shown in Figure 1. Each game admits two pure Nash equilibria: a Pareto-efficient equilibrium (A, A) and a Pareto-inferior equilibrium (B, B). As can be readily verified, g_{58} is a neighbor of g_{59} (swapping payoffs 2 and 3 of the column player in g_{58}), which in turn is a neighbor of g_{60} (swapping payoffs 1 and 2 of the column player in g_{59}). However, g_{58} is not a neighbor of g_{60}, and none of g_{58}, g_{59}, or g_{60} are neighbors of g_{70}. Despite these structural similarities, the four games fall into two level-k-based classes and three distinct empirical similarity classes.

To see why, consider first games g_{60} and g_{70}. Under the variant of the level-k model, all level-k types of both row and column players choose A in g_{60}, placing this game in the theoretical one-outcome level-k similarity class. The same holds for g_{70}, except that all level-k types play B, corresponding to the inefficient equilibrium. Hence, both belong to the same theoretical one-outcome level-k class. Empirically, however, their classifications differ: g_{60} falls into the one-outcome class $A1$, as play frequencies are concentrated on (A, A), while g_{70} breaks away to the four-outcome class $B2$, since observed frequencies are roughly balanced across actions for both players. The efficiency motive appears empirically important yet is not accounted for by the level-k classification. By contrast, in games g_{58} and g_{59}, level-k types of row and column players choose both A and B, implying that the two games belong to the theoretical four-outcome level-k similarity class. Empirically, though, they are assigned to the two-outcome class $A2$ (the one-sided dominance game class, with level-2 reaching the NE), as observed play is concentrated on the outcomes (A, A) and (A, B) since the row player's choice is concentrated on A, which is the level-1 action and which coincides with the action leading to the efficient NE. Note also that, although g_{59} is a neighbor of both g_{58} and g_{60}, there is continuity in empirical behavior only between g_{59} and g_{58}, not between g_{59} and g_{60}. Hence, they are in different empirical components.

The divergence between theoretical and empirical similarity classes typically arises when level-k predictions are neither efficient nor equitable. The case of g_{70} provides a clear example, as do g_{58} and g_{59}. Similar discrepancies also appear in prisoner's dilemma–type games.

The remainder of the paper is structured as follows. Section 2 introduces the 2×2 games that serve as the focus of our analysis. Section 3 presents the solution concepts and their corresponding decision rules used in the study. Section 4 provides formal definitions of game similarity (theoretical and empirical) and explores the similarity classes for some specific decision rules such as Nash equilibrium and level-k. Section 5 describes the design of our experiment, while Section 6 discusses the main results. Additional graphs, tables, proofs and formal definitions are provided in the Appendix.

2 Games and perspectives

Throughout the paper, we focus on two-player 2×2 games, where each player's payoffs are drawn from the set $\{1, 2, 3, 4\}$ without repetition.

Since there are 4! ways to arrange the payoffs for each player, the total number of possible games is $(4!)^2 = 576$. Let G denote the set of all such games. As is standard in game theory, two games are considered equivalent if one can be transformed into the other through a relabeling of players, a relabeling of actions, or any combination of these transformations. Grouping games into equivalence classes based on this criterion reduces the 576 games in G to 78 strategically distinct games, which we denote as G^*. Each equivalence class in G^* contains multiple payoff matrices that are equivalent through relabeling: 66 asymmetric games can be identified with eight equivalent games each, while 12 symmetric games have four equivalent games each. For simplicity, we refer to these equivalence classes as *games* in G^*. Section D in the Appendix provides a complete list of the representative games in G^*, numbered $1, \ldots, 78$. These numbers correspond to the node labels used in the graphs throughout the paper.

We define a *perspective* of a game as the choice problem a player faces, either as player 1 or player 2. In asymmetric games, there are two perspectives because the players face different choice problems, whereas in symmetric games, both players face the same choice problem, resulting in a single perspective. For example, of the four games in the Introduction, games g_{58}, g_{59} and g_{70} are asymmetric as players 1 and 2 face different choice problems, while game g_{60} is symmetric as both players face the same choice problem. In symmetric games, it can be verified that rewriting the bimatrix as if the column player were the row player, the resulting payoff matrix will be identical to the original. The 78 strategically distinct games in G^* therefore give rise to 144 perspectives: $66 \cdot 2 + 12 = 144$. We denote the set of all 144 perspectives as G^{**}. Formally, G^{**} is a set of equivalence classes of perspectives derived from the games in G, where each perspective, as viewed from the point of either player 1 or player 2, is equivalent.

3 Behavioral rules

We introduce minimal notation for our 2×2 games. Each game is represented by a tuple $g = (I, A, \pi)$, where $I = \{1, 2\}$ is the set of players, consisting of player 1 (the row player) and player

2 (the column player); $A = A_1 \times A_2$, where $A_i = \{a_{i,1}, a_{i,2}\}$, represents player i's action space for $i \in I$; and $\pi = (\pi_1, \pi_2)$, where $\pi_i : A \to \{1, 2, 3, 4\}$, is player i's payoff function.[10] We also consider mixed actions over A_i, with $\Delta = \Delta_1 \times \Delta_2$, where $\Delta_i \equiv \Delta(A_i)$ denotes the space of mixed actions available to player i.

In our analysis, we include a broad set of rules drawn from traditional game theory and over 30 years of behavioral game theory. Some concepts involve multiple actions (e.g., Nash Equilibrium or Pareto Efficiency) or mixed actions (e.g., Nash Equilibrium or Risk-Dominant Nash Equilibrium). To simplify the analysis while retaining these recommendations, we focus exclusively on pure actions. We define a **rule** for player i as a map $R_i : G \to 2^{A_i}$, where 2^{A_i} is the set of subsets of A_i, including the empty subset. For example, the concept Equal Split may recommend no actions for some games in G, resulting in an empty subset. When a concept recommends multiple pure actions, we include all those actions in R_i. Similarly, when the recommendation is a mixed action, we include all pure actions in the support of the mixed action. We denote the rule-profile used by both players as $R = (R_1, R_2)$, and will often refer to it simply as a rule.

The following is a list of the main game theoretical solution concepts and behavioral rules that we discuss and use to define the corresponding rules R_i for each player $i \in I$ in any game $g \in G$:

Best-response-based rules:

These are the main concepts from game theory based on strategic reasoning.

- **Not Strictly Dominated (NSD):** A pure action $a_i^{NSD} \in A_i$ is not strictly dominated if there does not exist another action which strictly dominates it.
- **Nash Equilibrium (NE):** An action profile $a^{NE} \in \Delta(A)$ is a Nash equilibrium if no player can achieve a greater payoff by unilaterally deviating from a^{NE}.
- **Risk-Dominant Nash Equilibrium (RDNE):** An action profile $a^{RDNE} \in \Delta(A)$ is a risk-dominant Nash equilibrium if a^{RDNE} is an NE and, when it is in pure strategies, it has the greatest *deviation loss* compared to other Nash equilibria. The deviation loss of a (Nash) action profile is the product of every player's payoff loss from deviating from that profile.
- **Rationalizability (RAT), Bernheim (1984), Pearce (1984):** An action profile $a^{RAT} \in A$ is rationalizable if it survives iterated elimination of strictly dominated actions.
- **Quantal Response Equilibrium (QRE), McKelvey and Palfrey (1995):** A (mixed) action profile $\alpha^{QRE(\lambda)} \in \Delta(A)$ is a quantal response equilibrium for some fixed $\lambda > 0$ if, for every $a_i \in A_i$, $\alpha_i^{QRE(\lambda)} = \exp\left(\lambda U(a_i, \alpha_{-i}^{QRE(\lambda)})\right) \Big/ \left(\sum_{a'_i \in A_i} \exp\left(\lambda U(a'_i, \alpha_{-i}^{QRE(\lambda)})\right)\right)$ where $U(a_i, \beta)$ is shorthand for the expected utility of player 1 playing the pure strategy a_i when player 2 plays $\beta \in \Delta(A_{-i})$. [11]

[10] Strictly speaking, in our class of games with payoffs in $\{1, 2, 3, 4\}$ without repetitions, the payoff functions π_i are bijections from A to $\{1, 2, 3, 4\}$.

[11] Unless stated otherwise, λ, a commonly known error, the same for all players, is selected to minimize the test

Level-k rules and variants:

- **Level-0 (L0):** The mixed action $a_i^{L0} = (1/2, 1/2) \in \Delta(A_i)$ for player i, which randomizes uniformly over i's actions. Such a player does not attend to any feature of the game.
- **Level-k (Lk), Nagel (1995):** A (mixed) action $a_i^{Lk} \in \Delta(A_i)$ is level-k for i if it is a best-response to the opponent playing a level-$(k-1)$ action.[12]
- **Level-$k(\alpha)$ (L$k(\alpha)$), Fudenberg and Liang (2019):** A (mixed) action $a_i^{Lk(\alpha)} \in \Delta(A_i)$ is level-$k(\alpha)$ for i if, after transforming all payoffs π_j into $\tilde{\pi_j} = \pi_j^{\alpha}$ for both players $j = 1, 2$, $a_i^{Lk(\alpha)}$ is a level-k action for i, where $0 < \alpha < 1$.[13] Notably, level-5(α) is equivalent to the NE action when it is unique and the L1(α) action otherwise.

Efficiency and/or equity-based rules:

For these rules, a player formulates an individual or collective goal of outcomes without strategic reasoning.

- **Pareto Efficiency (PE), Shubik (2012) [14]:** An action profile $a^{PE} \in A$ is Pareto efficient if every action profile that could improve one player's payoff compared to a^{PE} makes another player strictly worse off. The set of Pareto efficient profiles for a game $g \in G$ is denoted $PE(g)$.
- **Near-Equal Split (NES):** An action profile $a^{NES} \in A$ is a Near-Equal Split if $a^{NES} \in PE(g)$ and it minimizes the payoff difference between players among Pareto efficient outcomes. [15]
  - **Self-favoring NES (sNES):** An NES where the acting player has a weakly greater payoff than the opponent.
  - **Other-favoring NES (oNES):** An NES where the acting player has a weakly lower payoff than the opponent.
  - **Equal Split (ES):** An NES with zero payoff difference between players (may be empty for some games).
- **Max-Max (MM):** An action $a_i^{MM} \in A_i$ is Max-Max for i if it can lead to the outcome with the highest payoff for i.
- **Soc-Max (SM):** An action profile $a^{SM} \in A$ maximizes the sum of payoffs, $\sum_{i \in I} \pi_i(a)$, over all $a \in A$.

MSE from 10-fold cross-validation of a sample. This ensures both that QRE is not an overfitting model whilst also giving it the best chance of being a well-fitting model.

[12] In our 2×2 games, we consider k only up to 5 due to cyclicality of the rule. For instance, level-6 is the same as level-2, level-7 is level-3, and so on. A level-1 player only attends to own payoffs.

[13] This transformation ensures that $Lk(\alpha)$ rules yield unique pure actions for all $k \geq 1$ in 2×2 games. Both players' payoffs are transformed in the calculation. Any α <1 yields the same action. This rule serves as a tie-breaker for our games when the payoff sum of both actions is $5 = 4 + 1$ and $3 + 2$, respectively.

[14] Shubik (2012) sorted the same 2x2 games as ours according to the Pareto frontier.

[15] This concept relates to equity principles, social preferences, or team-reasoning, see literature mentioned in the introduction.

Hybrid rules:

- **Pareto-Dominant Nash Equilibrium (PDNE), Fudenberg and Liang (2019):** An action profile $a^{PDNE} \in \Delta(A)$ is a Pareto-dominant Nash equilibrium if it is an NE and Pareto dominates every other NE. By definition, any game with a unique NE automatically satisfies this criterion. In our 2×2 games, this means that in games with a unique equilibrium, the equilibrium strategy is chosen. In coordination games with Pareto-ranked equilibria, the action leading to the Pareto optimal outcome $(4, 4)$ is chosen. In other coordination games where there are two PONE as in battle of sexes games, and in Matching Pennies-type games, players mix between the two actions.

As a first step, we use these game-theoretical and behavioral rules to define theoretical and behavioral similarity classes for the games, assuming all subjects use one rule or class of rules. In the results sections, we estimate empirical similarity classes based on aggregate behavior and compare them to the theoretical similarity classes. Finally, we identify and analyze each subject's rules based on their choices across the 144 perspectives.

4 Defining similarity classes of games

In this section, we introduce a notion of similarity that groups games according to the continuity a rule maintains through the space of games. For that notion of similarity to be coherent, we need to formalize how G^* simplifies G into equivalence classes and define a topology that allows us to measure the proximity of games to one another in G.

Thus, we first construct a metric space based on proximity of games that results in one similarity class, ordering all games according to a minimal distance between two neighboring games, see also Shepard (1987) for a similarity-based metric space of psychological objects. Each rule or class of rules mentioned in the previous section provide the feature sets (a la Tversky, 1977), according to which a subject attends to and chooses an action. Combining the metric space of games and a particular rule creates a subgrouping, resulting in similarity classes of the entire set of games.

4.1 Preliminaries

Any given 2×2 game g, can be identify by its associated payoff matrix, defined by:

$$\Pi = \begin{bmatrix} a, e & b, f \\ c, g & d, h \end{bmatrix},$$

which can also be written as the tuple $(a, b, c, d, e, f, g, h) \in \{1, 2, 3, 4\}^8 \subset \mathbb{R}^8$, and where it is understood that $a = \pi_1(a_{1,1}, a_{2,1}), \ldots, d = \pi_1(a_{1,2}, a_{2,2})$ and $e = \pi_2(a_{1,1}, a_{2,1}), \ldots, h = \pi_2(a_{1,2}, a_{2,2})$.[16]

Moreover, as mentioned before, there are a total of 7 more games represented by their respective

[16]There is a slight abuse of notation here and the following paragraph in that g and a are also used as payoff entries in Π. In general, g denotes a game and a denotes an action profile unless otherwise noted.

payoff matrix, that are fundamentally the same as the game represented by Π, namely:

$$\begin{bmatrix} c,g & d,h \\ a,e & b,f \end{bmatrix}, \begin{bmatrix} b,f & a,e \\ d,h & c,g \end{bmatrix}, \begin{bmatrix} e,a & g,c \\ f,b & h,d \end{bmatrix}, \begin{bmatrix} d,h & c,g \\ b,f & a,e \end{bmatrix}, \begin{bmatrix} g,c & e,a \\ h,d & f,b \end{bmatrix}, \begin{bmatrix} f,b & h,d \\ e,a & g,c \end{bmatrix}, \begin{bmatrix} h,d & f,b \\ g,c & e,a \end{bmatrix}$$

All of these can be obtained from Π by either switching rows (i.e., relabeling 1's actions), switching columns (relabeling 2's actions), transposing and switching the off-diagonal entries (relabeling player 1 as 2 and player 2 as 1), or some combination of these operations. And so, although they represent different tuples in $\mathbb{R}^8$, they all represent the same strategic interaction up to relabeling of actions and players. Formally, the operations of relabeling actions and players and their combinations are linear maps $\psi : \mathbb{R}^8 \to \mathbb{R}^8$ that include the identity map on $\mathbb{R}^8$, so that the set of all such symmetry operations Ψ, contains exactly eight different linear maps, one being the identity map and the remaining seven being the ones mapping Π to one of the seven transformed payoff matrices above. Following (36) and (29), we refer to the maps ψ as **symmetry operations**, and we say any two games associated through such a map are **equivalent**. In particular, all games above are equivalent to the game represented by the payoff matrix Π.

This notion of equivalence allows us (and others in the literature) to reduce the number of games from 576 (in G) to 78 (in G^*). This reduction can be done by brute force using graph theoretic methods. As is standard, we use $\mathbf{\Gamma} = (\Gamma, E)$ to denote a graph with vertices Γ and edges $E \subseteq \Gamma \times \Gamma$. We define $E(A)$ as the edges induced by the adjacency matrix A so that $(g, g') \in E(A)$ implies that $A(g, g') \neq 0$. Finally, we say that $\mathbf{\Gamma}_c$ is a component of $\mathbf{\Gamma}$ if it is a connected subgraph of $\mathbf{\Gamma}$ that is not part of any larger connected subgraph of $\mathbf{\Gamma}$.[17]. If one defines the 576×576 adjacency matrix A_Ψ so that for $g \neq g'$

$$A_\Psi(g, g') = \mathbb{1}\{\exists \psi \in \Psi : g = \psi g'\},$$

and $A_\Psi = 0$ otherwise, we show in the Appendix (see figure A1) that there are 78 components of the graph $(G, E(A_\Psi))$, meaning that there are only 78 games that are unique up to symmetry operations.[18]

Our notion of two games being neighbors is directly taken from Robinson and Goforth (2006):

Definition 1 (Neighboring games)**.** *Let $g, g' \in G$. We say that g' is a* neighbor *of g (or $g' \in N(g)$) if g' can be obtained from g by swapping two entries in the payoff matrix of g for one of the players only, and the two entries differ 1. $N(g) \subset G$ is the set of all neighbors of g, including g itself.*

For example, $\begin{bmatrix} 4,4 & 2,1 \\ 1,2 & 3,3 \end{bmatrix}$ is a neighbor of $\begin{bmatrix} 4,4 & 1,1 \\ 2,2 & 3,3 \end{bmatrix}$, but is not a neighbor of $\begin{bmatrix} 4,4 & 1,2 \\ 2,1 & 3,3 \end{bmatrix}$.

Notice that $N(g)$ coincides with the set of all games in G with Euclidean distance $d \leq \sqrt{2}$ from g. The definition above induces a topology on the space G that we refer to as the Robinson-Goforth

[17] A graph (Γ, E) is connected if for any $g, g' \in \Gamma$, there exists a path from g to g', that is, a sequence $(g_1, \ldots, g_n)$ such that $g_1 = g, g_n = g'$ and $(g_k, g_{k+1}) \in E$. A graph (Γ', E') is a subgraph of (Γ, E) if $\Gamma' \subseteq \Gamma$ and $E' \subseteq E$

[18] $\mathbb{1}\{\cdot\}$ is short-hand for the indicator function.

topology, which is studied in great detail in (41). It can be represented by a graph, $(G, E(A_N))$ where A_N is a 576×576 adjacency matrix such that for $g \neq g'$: a path of neighboring games can join g and a game equivalent to g'.[19]

4.2 Similarity of games

We can now define our notion of similarity, which follows Germano (2006) and defines two games $g, g' \in G$ as similar according to a rule R if a path of neighboring games can join g and a game equivalent to g' in G, where the rule R always prescribes the same action or set of actions for all games along the path. This ensures that g and a game equivalent to g' belong to the same connected component of games in G. Formally:

Definition 2 (Similarity of games)**.** *Let R be a rule. We say that g is* similar *to g' according to R (written as $g \sim_R g'$) if and only if there exists a path $(g_1, g_2, \ldots, g_n)$ such that $g = g_1$, $g' = \psi g_n$, for some symmetry operation $\psi \in \Psi$, and, for any two consecutive games $g_\nu, g_{\nu+1}$, along the path, $g_{\nu+1} \in N(g_\nu)$ and $R(g_\nu) = R(g_{\nu+1})$, for $\nu = 1, \ldots, n-1$.*

If we are given a finite set of rules $\mathcal{R}$, then we say $g \sim_{\mathcal{R}} g'$ if the same conditions hold as with one rule, but where the condition $R(g_\nu) = R(g_{\nu+1})$ now holds for all $R \in \mathcal{R}$.

It can be checked that, for any rule R, the similarity relation $\sim_R$ uniquely partitions the set of games G into a finite number of components, which we refer to as the **similarity classes** in G relative to R. By definition of the symmetry operations, if we consider the smaller set G^* instead of G, then $\sim_R$ also uniquely partitions G^* into the same number of components as in G. Hence, the similarity classes of G are in a one-to-one relation with the similarity classes of G^*. It is important to emphasize that two equivalent games need not be neighbors and that, in general, two equivalent games need not be connected by a path of neighboring games along which a given rule is constant. Nonetheless, the fact that there is a symmetry operation linking the two games, ensures that they belong to the same similarity class.

4.3 Computation of similarity classes of games

Computing the similarity classes for the rules in Section 3 is feasible by combining adjacency matrices through matrix operations, as we show in the following theorem. Let $A_R(g, g') = \mathbb{1}\{R(g) = R(g')\}$ and denote the Hadamard (or element-by-element) product of two matrices A and B by $A \odot B$ with $(A \odot B)(g, g') = A(g, g')B(g, g')$. Finally, we say that rule R is *relabeling invariant* if $R(g) = R(g')$ implies that for any $\psi \in \Psi$, $R(\psi g) = R(\psi g')$.[20]

Theorem 1 (Computation of Similarity Classes)**.** *Let R be relabeling invariant. Then, $g \sim_R g'$ if and only if g and g' belong to the same component of $(G, E(A_\Psi + (1 - A_\Psi) \odot A_N \odot A_R))$. Thus,*

[19] Robinson and Goforth (41) not only show that the graph $(G, E(A_N))$ cannot be embedded in a plane (unlike the Periodic Table of Elements, which has a Euler characteristic of 2), but that it requires a 37-holed surface to represent all the links without crossing (as it has Euler characteristic -72). The complexity of this space arises from the interaction of the symmetry operations and neighboring games on a space of games with payoffs 1, 2, 3, and 4, which includes symmetric games, all of which together make up the topology of the graph.

[20] It is easy to show that all the decision rules in Section 3 satisfy this property.

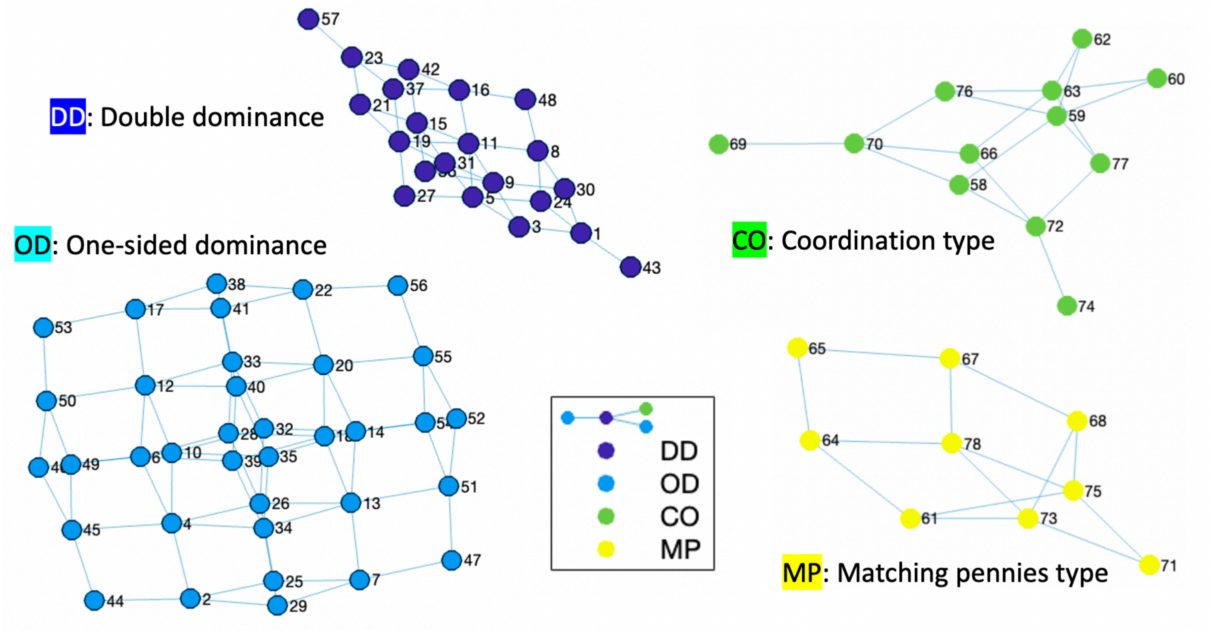


Figure 2: Graph of the similarity classes in G^* implied by the NSD rule. Each node is a game in G^* and a link is drawn between a pair of games if: (i) for some representation of the pair of games in G, both games are neighbors and (ii) both games belong to the same similarity class.

the similarity classes according to R over G correspond precisely to the components of $(G, E(A_\Psi + (1 - A_\Psi) \odot A_N \odot A_R))$.

The proof is in the Appendix. The theorem makes operational the idea that similar games are connected either by the symmetry operations (thus being in the same component defined by A_Ψ) or through the neighborhood topology by a constant rule recommendation given R (thus being in the same component defined by $((1 - A_\Psi) \odot A_N \odot A_R)$. The theorem states that these components characterize the similarity classes for the given rule R.

4.4 Examples of theoretical similarity classes

We now present the similarity classes for some key rules. To simplify the figures, we visualize only the similarity classes over the 78 games in G^*. We refer to the table in Section D of the Appendix for a list of all the games in G^*.

The first is a very basic and well-known partition (see also Rapoport, Guyer, and Gordon, 1976; Bruns, 2015; Biggar and Shames, 2023) that serves as a basis of analysis throughout the paper. The coloring for the four types of games obtained here from the four similarity classes are used for all the figures of similarity classes that follow.

Strict dominance similarity classes. If $R : G \to 2^A$ is the rule choosing actions that are Not Strictly Dominated (NSD), then it partitions the games in G (and hence G^*) into four similarity classes as follows (see Figure 2):

- **Double-sided dominance games (DD):** Games with a unique pure Nash equilibrium,

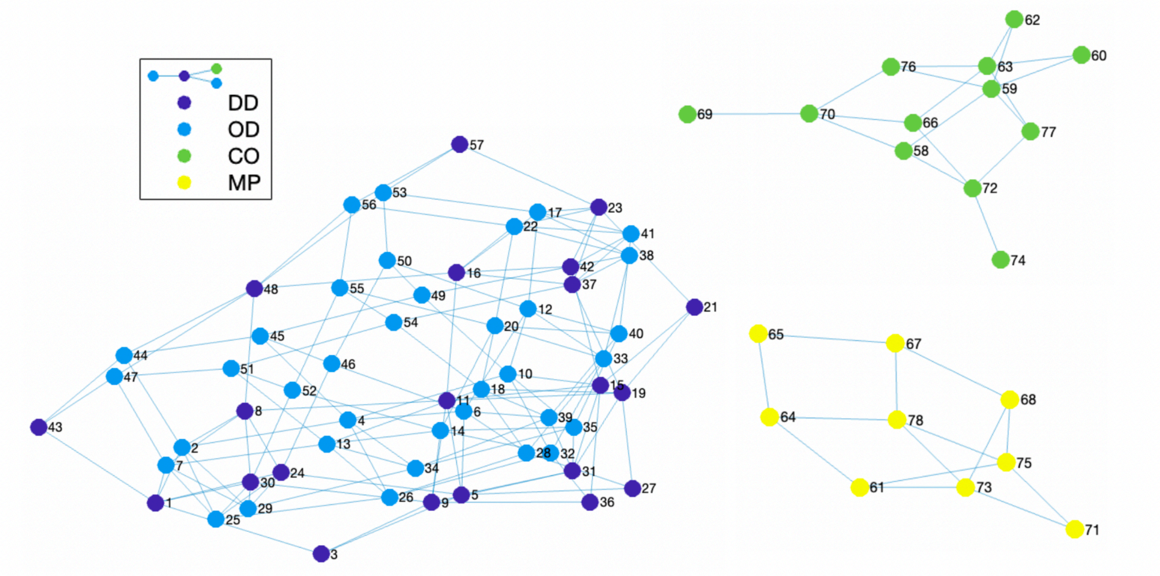


Figure 3: Graph of the similarity classes in G^* implied by Nash equilibrium. Each node is a game in G^* and a link is drawn between a pair of games if: (i) for some representation of the pair of games in G, both games are neighbors and (ii) both games belong to the same similarity class.

where both players have a strictly dominated action (dark blue nodes, 21 games).

- **One-sided dominance games (OD):** Games with a unique pure Nash equilibrium, where only one player has a strictly dominated action (light blue nodes, 36 games).
- **Coordination-type games (CO):** Games with two pure NE; neither player has a strictly dominated action (green nodes; 12 games).
- **Matching pennies-type games (MP):** Games with a unique mixed NE and no pure NE; neither player has a strictly dominated action (yellow nodes; 9 games).

We refer to these four classes as the *strict dominance* similarity classes.[21] The reason that the games in CO and the ones in MP are separated, although no actions are dominated in both classes is that these two classes (yellow and green nodes) is due to the topology of neighboring games, since one has to pass through games in DD or OD (dark and light blue nodes) to move from a game in one class (say, CO) to a game in the other (say, MP) (see Figures A3 and A4 in the Appendix). This means that there cannot be paths connecting games from CO to ones in MP, where the NSD rule is continuous. ■

The next similarity classes are the ones associated with the Nash Equilibrium rule.

Nash Equilibrium similarity classes. If $R : G \to 2^A$ is Nash Equilibrium rule (NE), then it partitions the games in G (and hence G^*) into three similarity classes as follows (see Figure 3):

- **Games with a unique pure NE (DD ∪ OD):** These games are all dominance solvable

[21] Biggar and Shames (2023) obtain the same classification based on the notion of *response graphs* without using the neighborhood structure.

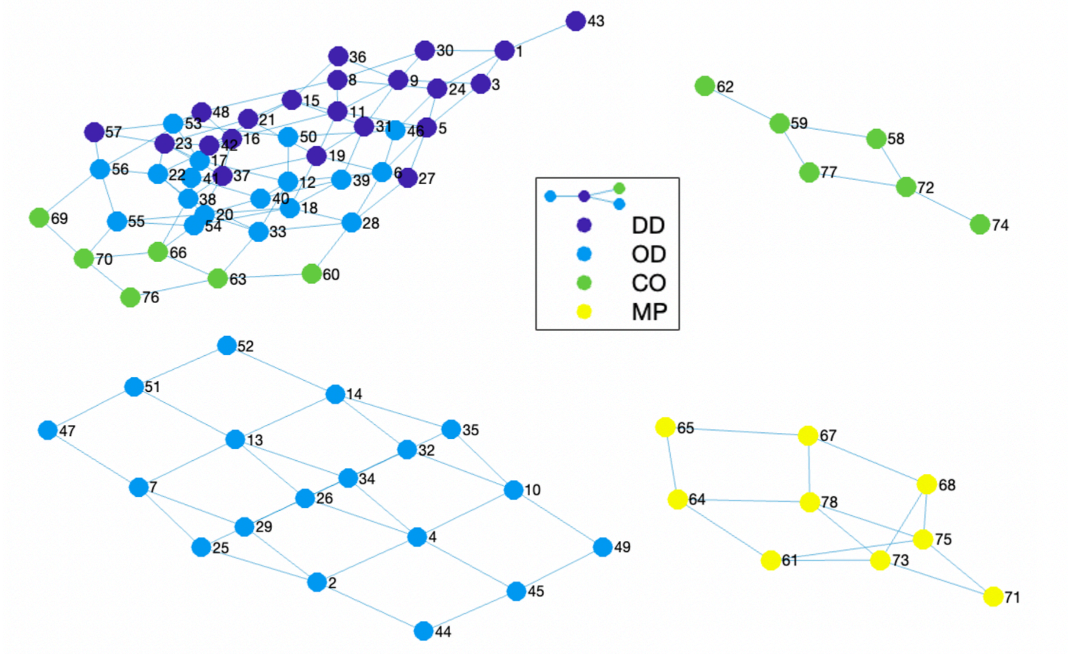


Figure 4: Graph of the similarity classes in G^* implied by level-$k(\alpha)$ rules, $k = 1, \dots, 5$. Each node is a game in G^* and a link is drawn between a pair of games if: (i) for some representation of the pair of games in G, both games are neighbors and (ii) both games belong to the same similarity class.

and thus contain all DD and OD games (dark blue and light blue nodes; 57 games);

- **Games with two pure NE (CO):** These are all the coordination type games (green nodes; 12 games);
- **Games with zero pure NE (MP):** These are all the matching pennies type games (yellow nodes; 9 games).

The same similarity classes arise if one considers the rule based on Rationalizability (RAT). Again, the games in CO and the ones in MP are separated although all actions are always rationalizable or in the support of a Nash equilibrium. The reason is the same as with the NSD rule, since it is necessary to passs through games in DD or OD to go from a game in CO to a game in MP (see Figures A3 and A4 in the Appendix). ■

Next we discuss the similarity classes for the level-$k(\alpha)$ rule that play a key role in our experimental results.

Level-$k(\alpha)$ similarity classes. If we use rules based on level-$k(\alpha)$ (L$k(\alpha)$, $k = 1, \dots, 5$), we obtain different similarity classes from Nash. Taking all L$k(\alpha)$ rules for $k = 1, \dots, 5$, as a set $\mathcal{R}$ of rules, leads to four similarity classes being distinguished, see Figure 4, namely:

- **One-outcome games (DD ∪ OD1 ∪ CO1):** These are games where all the L$k(\alpha)$ actions of both players span only One-outcome.[22] In particular, the L$k(\alpha)$ actions reach a pure Nash

[22] We say the L$k(\alpha)$ actions of both players **span** ℓ **outcomes**, if the number of distinct profiles contained in the product of the set of L$k(\alpha)$ actions for both players is ℓ. Formally, $\#\{a_1^{L1(\alpha)}, a_1^{L2(\alpha)}, \dots\} \times \{a_2^{L1(\alpha)}, a_2^{L2(\alpha)}, \dots\} = \ell$.

equilibrium in just one step (L1(α) with all higher levels choosing that same action). These games include all the DD games as well as some OD and some CO games, henceforth referred to as OD1 and CO1, respectively (component of dark blue, light blue, and green nodes in Figure 4; 45 games).[23]

- **Two-outcomes games (OD2):** These are games where L$k(\alpha)$ actions span Two-outcomes: player 1 always chooses the dominant action, thus L1(α) and player 2 chooses L1(α) or L2(α) since two levels $k = 2$ are needed to reach the unique Nash equilibrium. All higher Lk(α) choose the same action as L2(α). These games consist only of a subset of OD games (those which are not in the One-outcome class), with L1(α) strategy combination not forming a NE, henceforth referred to as OD2 (component of light blue nodes in Figure 4; 18 games).
- **Four-outcome games with two pure NE (CO2):** These are games where L$k(\alpha)$ actions span all Four-outcome and hence also the two pure strategy Nash equilibria. These games consist of a subset of all CO games (not in the One-outcome class), henceforth referred to as CO2 (component of light green nodes in Figure 4; 6 games).
- **Four-outcome games with zero pure NE (MP):** These are games where L$k(\alpha)$ cycles over all outcomes, thus spanning Four-outcome. This coincides precisely with the class of MP games defined above for the NE-concept (component of yellow nodes in Figure 4; 9 games).

How these similarity classes are connected on the graph of G is shown in the Appendix (Figure 4 and A7). The separation of different classes is based on the number of outcomes reached by the players playing the different levels k. This naturally separates the One-, Two-, and Four-outcome games. However, there are two separate Four-outcome classes consisting of the coordination-type games (CO2) on the one hand and matching pennies-type games (MP) on the other. These two classes cannot be connected due to the neighborhood structure of the games in G. Any path between a game in CO2 and a game in MP has to pass through either the component of OD or DD games with 1 or 2 outcomes. ■

The similarity classes obtained for the four rules discussed above form the basis for characterizing the empirical similarity classes obtained in Section 5. These systems are quite different from each other. However, they also share some common features. They agree on separating the MP games from the rest, confirmed by the empirical features. All DD games belong to one class. Coordination games and OD games, respectively, can form a separate class or (partially) join the DD class. In the Appendix, we discuss other similarity classes, such as the ones obtained from Risk Dominant Nash Equilibrium or the rules based on Pareto efficiency (PE) or Near Equal Split (NES).

4.5 Empirical similarity of games

To conclude this section, we extend the notion of similarity to study the empirical behavior of subjects, yielding the empirical similarity classes of games. For this, define an **empirical rule**

[23] The Table of Games in G^* at the end of the Appendix lists the L$k(\alpha)$ type of game – whether DD, OD1, CO1 etc. – for each of the 78 games in G^*.

with tolerance, $\hat{R}_\delta : G \to 2^\Delta$, as a map assigning confidence intervals around the frequency of play in game g for both players (1 and 2). These intervals can be derived from bootstrapping or the binomial test. Unlike the original definition, where the condition required exact equality in play behavior, we introduce a tolerance parameter, $\delta > 0$, to allow for statistical variability. Formally, the test evaluates whether the difference in behavior between games lies within the interval $[-\delta, \delta]$ at a fixed confidence level (e.g., 99%).

This approach aligns with a statistical equivalence testing framework also known as the Two One-Sided Tests (TOST) procedure (Schuirmann, 1987). Intuitively, TOST checks whether the average difference in play behavior between two games is small enough to be considered equivalent, rather than requiring the difference to be strictly zero. This relaxation ensures the method is robust to sampling variability. Nonetheless, when $\delta = 0$, it is easy to see that TOST is equivalent to the standard two-sided paired differences test.

Definition 3 (Empirical similarity of games)**.** *Let $\hat{R}_\delta : G \to 2^\Delta$ be an empirical rule with tolerance $\delta > 0$. We say that g is* empirically similar with tolerance δ *to g' according to $\hat{R}_\delta$ (written as $g \sim_{\hat{R}_\delta} g'$) if and only if there exists a path $(g_1, g_2, \dots, g_n)$ such that:*

- *$g = g_1$, $g' = \psi g_n$, for some symmetry operation $\psi \in \Psi$;*
- *For any two consecutive games $g_\nu, g_{\nu+1}$ along the path, $g_{\nu+1} \in N(g_\nu)$, and the bootstrapped confidence interval for the average difference satisfies:*

$$\frac{1}{|S|}\sum_{s=1}^{|S|}(A_{sg_\nu} - A_{sg_{\nu+1}}) \in [-\delta, \delta],$$

for $\nu = 1, \dots, n-1$, where $A_{s\gamma}$ is an indicator function for whether subject s plays the top row action (labeled as A) in game γ and $|S|$ is the sample size.

From the above definition, we obtain **empirical similarity classes with tolerance**, which partition the space G (or equivalently G^*) into a finite number of components. These partitions depend on the empirically observed behavior of subjects and the tolerance parameter δ.

The use of δ allows for a flexible classification, enabling the comparison of behaviors across games while accounting for statistical uncertainty. In the results section, we compare the theoretical similarity classes from game-theoretic solution concepts with the empirical ones derived from experimental data.

5 Experimental design

We recruited 450 subjects from Prolific, with ethical approval from CIREP-UPF. To control for regional variations, the sample was restricted to participants based in the United Kingdom. The majority of subjects were women (295 out of 450; 65.56%) and men made up the remainder (153 out of 450; 34.00%). Participants were predominantly aged 25-44, with 30.67% (138 out of 450)

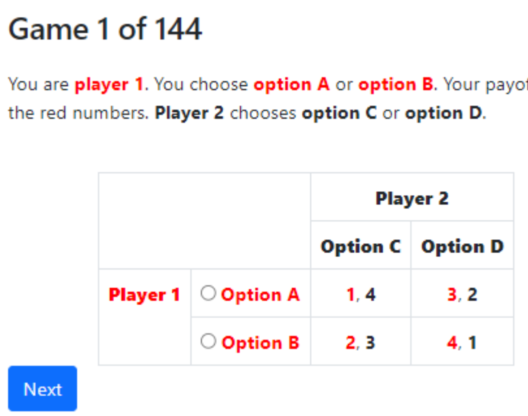


Figure 5: A screenshot of the interface for a decision problem in Stage 1 of the experiment.

in the 25-34 age group and 30.44% (137 out of 450) in the 35-44 age group. Regarding education, 24.44% (110 out of 450) had completed some college education, and 28.89% (130 out of 450) held a four-year degree.

The experiment was programmed using oTree (Chen, Schonger, and Wickens, 2016), and each subject participated in a standalone version of the experiment. The experiment consisted of three stages: (1) a 2×2 game-playing stage, (2) a risk-elicitation task, and (3) an exit questionnaire describing decision processes.

5.1 Stage 1: Decisions for 144 Perspectives

Each game was constructed using only four distinct payoffs—1, 2, 3 and 4—for each player, without replacement, as described in the preceding sections.[24] Each subject always played in the position of the row player, choosing between the top and bottom rows (labeled as option A and option B in the interface) for all 144 perspectives. This required subjects to play both roles (row and column player) across all 78 games (66 asymmetric games and 12 symmetric games). A screenshot of the decision interface is provided in Figure 5.

This design allowed us to reduce the required number of subjects while simplifying the selection and ordering of games for each participant. In a pretest, we evaluated whether playing all games (144 perspectives) instead of a smaller subset (44 perspectives) introduced behavioral changes due to potential boredom. No significant differences in behavior were observed between the two conditions (see Section C of the Appendix for details on the comparison between the 44-game and all-games experiments).

No feedback was provided between rounds to prevent learning from other subjects. To mitigate fatigue, subjects were given a 30-second break after completing their 48th and 96th games. These breaks were not announced in advance; subjects only became aware of them upon encountering the first break after the 48th game. Additionally, since no time limit was imposed, subjects were

[24]Martin Shubik (2012) did a pilot with a similar design at the Stony Brook Game Theory festival. He proposed the experiment to Shyam Sunder who told Nagel about it more than 10 years later.

free to take further breaks as needed.

A sequence of 144 Perspectives

Many games are theoretically similar, as discussed in the previous section. To avoid random bunching of theoretically similar games, which might cause subjects to expect future games to resemble previous ones, we introduced structure in the ordering of games instead of generating sequences purely randomly. For example, if subjects encountered game g_1 and g_2 consecutively, both featuring the outcome $(4, 4)$ (see Section D of the Appendix), they might develop an expectation that future games would also lead to outcomes where both players receive the maximum payoff. This could bias their behavior, such as favoring the near-equal split profile.

To address this, we assigned a numerical code to each perspective in G^{**} based on its payoff matrix $\Pi = (a, b, c, d, e, f, g, h)$. The code was defined as "abcdefgh," where each letter represents a payoff value in the matrix. For instance, the Prisoner's Dilemma with $\Pi = (4, 2, 3, 1, 1, 2, 3, 4)$ is assigned the code "42311234" (see perspective 57 in Section D). We then ordered these numerical codes in increasing order and grouped them into 16 quantile bins, each containing 9 perspectives ($16 \times 9 = 144$ perspectives).

To test for game order effects, we generated 15 different sequences of the 144 perspectives using the following procedure:

1. Randomly generate an order of the 16 quantile bins (e.g., 1, 5, 9, 14, 4, 16, 10, 12, 6, 8, 11, 2, 3, 7, 15, 13).
2. Randomly select one perspective from each quantile bin without replacement, following the order generated in Step 1 (e.g., the first perspective comes from bin 1, the second from bin 5, the third from bin 9, and so on), and append it to the sequence.
3. Repeat Steps 1 and 2 until the sequence includes all 144 perspectives. [25]

This method minimizes the likelihood of random bunching, as each block of 16 randomly chosen perspectives includes one perspective from each bin. In each sequence, the randomization of labels is performed as follows: for every perspective, we re-label either player 1's actions or player 2's actions using symmetry operations. Importantly, we do not re-label the players' positions, as this would alter the perspective.

Finally, one of the 15 treatment sequences was randomly assigned to each of the 450 subjects, with 30 subjects per sequence. While previous experiments (e.g., Fudenberg and Liang, 2019) randomized sequences per subject, we opted for this design to test whether the order within a sequence has a statistically significant effect.

[25] Subjects play the perspectives from G^{**} rather than all games in G because behavior is assumed to be invariant to relabeling. However, this assumption is not taken for granted; we control for this by randomizing the labels of games and balancing their assignment.

5.2 Stage 2: Risk-elicitation

In stage 2 of the experiment, we elicited subjects' risk preferences using the risk-taking item from the Global Preferences Survey (Falk et al., 2018). Subjects first completed the original item as designed for their country in the Global Preferences Survey. Following this, they completed a modified version of the same item, where the payoffs were re-scaled to match the payoff scale of the games played in stage 1. The positive affine transformation adjusted the maximum value of the lotteries in the task to correspond to the monetary value of receiving 4 points in the games, while the minimum value was adjusted to match the monetary value of receiving 1 point. All values were presented in the currency of the subject's country to ensure consistency and comprehension. Analysis of the results from this stage is forthcoming.

5.3 Stage 3: Exit questionnaire

In the exit questionnaire, we asked subjects to describe in their own words how they came to make decisions in the games. We also asked subjects questions on a Likert scale about how they perceived their strategy across stage 1 and information about their demographics (e.g., age, education, gender identity, to name a few).

5.4 Payment

The average completion time for the entire experiment was approximately 20 minutes, with subjects taking an average of 16.3 minutes (median: 10.8 minutes) to complete the 144 game decisions in stage 1. On a per-game basis, the average decision time was 6.8 seconds (median: 4.5 seconds).

Subjects were compensated with a show-up fee of 1.70 GBP upon completing the experiment, along with a performance-based payment of 2.00 GBP per point, calculated as the average number of points earned across three randomly selected games. This payment structure allowed subjects to earn between 2.00 GBP and 8.00 GBP, with an average total payment of 7.02 GBP (1.70 GBP + 2.00 GBP per point × 2.67 average points).

No feedback was provided between rounds, and subjects were randomly matched with a new opponent in each round. To mitigate fatigue, participants were given forced 30-second breaks after completing their 48th and 96th games.

6 Results

In this section, we first present the empirical similarity classes based on aggregated behavior and compare them with the theoretical classes mentioned above. We then analyze whether the modal behavior of the individual games can be explained by the rules suggested by the empirical similarity classes. Furthermore, we classify the behavior of each subject according to the behavioral rules. The level-$k(\alpha)$ and NES rules will drive most of the results. Additionally, to better understand individual behavior, we further study empirical similarity classes for subgroups of subjects' behavioral types.

6.1 Empirical similarity classes

As a first step, we aggregate all data for each game, separately for each player role, to derive the empirical similarity classes. The **empirical rule** $\hat{R}$ is estimated using the Two One-Sided Test (TOST) procedure to determine whether the average difference in choice frequencies between two games falls within a threshold value δ. For the remainder of the paper, we set $\delta = 5\%$. In the Appendix, we report results for other thresholds, including $\delta = 0\%$, $\delta = 1\%$, and $\delta = 2\%$.

The resulting similarity classes consist of eight distinct components, as shown in Figure 6 and also Figure 7 that shows the relative frequencies for each game. None of the theoretical or behavioral rule-based similarity classes *alone* described in Section 3 correspond exactly to these empirical similarity classes. However, the empirical similarity classes appear to be primarily characterized by the similarity classes corresponding to the L$k(\alpha)$ rules, and most exceptions seem to be driven by tensions with near-equal split outcomes.

A. Empirical similarity classes: Level-$k(\alpha)$ related components

A clear connection to the four theoretical L$k(\alpha)$ similarity classes (see Figure 4) is evident. The three main components that emerge correspond to the one-outcome, two-outcome, and four-outcome MP games, respectively. The four-outcome CO2 component reduces to a component of just 2 games out of the original 6. Beyond these four components, four new ones emerge, which we interpret as breaking away from the four theoretical L$k(\alpha)$ classes due to discontinuities in the aggregate behavior of neighboring games. This occurs essentially when the strategy profile predicted by L1(α) does not align with a near-equal split outcome (NES). We start with the four L$k(\alpha)$ related components and discuss the departures from the theoretical L$k(\alpha)$ classes in terms of games that remain, are added, or removed.

A1. One-outcome games (36 games of which DD: 20, OD1: 15, CO1: 1) **Features:** For both row and column players, all L$k(\alpha)$ actions lead to the same (pure) Nash equilibrium that is also a NES. **Exceptions:** $g_{17}, g_{37}, g_{48}, g_{53}$.

Relation to L$k(\alpha)$ one-outcome class: 36 remain, 9 removed, 0 added.

- **36 games remaining:** Of the 45 games in the theoretical one-outcome L$k(\alpha)$ class, 36 remain with relative frequencies higher than 70% of L1(α) actions (see Figure 7, upper-right corner). In all but 4 of these games, the L1(α) lead to a Nash equilibrium that is a NES; for the remaining 4 (g_{48} DD game, and g_{17}, g_{37}, g_{53} OD1 games), the Nash equilibrium is not a NES though it is Pareto efficient. Finally, of the 6 CO1 games, only g_{60} remains in this component.
- **9 games removed:** Nine games break away from the theoretical one-outcome L$k(\alpha)$ class where the L1(α) actions form a NE but do not constitute a NES. Game g_{63} is an

exception to this, as it has two outcomes, where the Nash equilibrium is NES, resulting in more heterogeneity not observed in the one-outcome class. These nine games form three new subclasses described in B1-B3 below, which belong to the class of PD, Stag hunt, and BoS games.

Representative game: $g_3 = \begin{bmatrix} 4,4 & 3,2 \\ 2,3 & 1,1 \end{bmatrix}$

A2. Two-outcome games (22 games of which OD2: 18, CO2: 3, MP: 1) **Features:** For one player, all L$k(\alpha)$ actions, leads to a (pure) Nash equilibrium, but for the other player the NE is reached with L2(α) action. **Exceptions:** g_{65} from MP, g_{74} from CO2.

Relation to L$k(\alpha)$ two-outcome class: 18 remain, 0 removed, 4 added.

- **18 games remaining:** All 18 games in the two-outcome L$k(\alpha)$ class remain. These are all the OD2 games. Figure 7 with the A2 component shows that subjects play the dominant strategy more than 80%. However, the opponent is divided between the two actions, with a range of 35% to 70% in different games, thus producing mainly two outcomes.

- **4 games added:** Four games are added to the two-outcome L$k(\alpha)$ class. Game g_{65} is from the L$k(\alpha)$ MP class and the three games g_{58}, g_{59}, g_{74} are from the four-outcome CO2 one. In the first three games, at least one player has an action with potential payoffs of 4 and 2, and in the Chicken game, g_{74}, the L1(α) action leads to a Nash equilibrium and NES making it attractive and resulting in a behavioral discontinuity from neighboring games in the L$k(\alpha)$ CO2 class, but creating a continuity with neighboring OD2 game g_{47} (see Figure 7, lower right corner).

Representative game: $g_7 = \begin{bmatrix} 4,3 & 3,4 \\ 2,2 & 1,1 \end{bmatrix}$

A3. Four-outcome CO2 games (2 games both CO2) **Features:** For both players the L1(α) action leads to a (pure) Nash equilibrium that is near-equal split but is different for the two players.

Relation to L$k(\alpha)$ four-outcome CO2 class: 2 remain, 4 removed, 0 added.

- **2 games remaining:** Of the 6 games originally in the CO2 class, only g_{72} and g_{77} remain together and appear as 4-outcome games, (see Figure 7, middle part). These are the only games of the CO2 class in which both players L1$S(alpha)$ actions leads to potential payoffs 3 or 2. which neither leads to a NE nor NES outcome, and spreads the frequencies so that it behaves empirically as a four-outcome game. The presence of two NES and the fact that the L1(α) actions have potential payoffs of just 3 and 2 spreads the frequencies so that it behaves empirically as a four-outcome game.

- **4 games removed:** three games g_{58}, g_{59}, g_{74} go to the two-outcome component A2 mentioned above; g_{62} forms its own component B4 (see below).

Representative game: $g_{77} = \begin{bmatrix} 4,3 & 1,1 \\ 2,2 & 3,4 \end{bmatrix}$

A4. Four-outcome MP games (8 games all MP) **Features:** For both players, the $L1(\alpha)$ action does not lead to a (pure) Nash equilibrium.

Relation to L$k(\alpha)$ four-outcome MP class: 8 remain, 1 removed, 0 added

- **8 games remaining:** Of the 9 games originally in the MP class, 8 remain.
- **1 game removed:** Game g_{65} is removed and goes to the two-outcome component A2 mentioned above. It is the only MP game where both players have an action with potential payoffs of 4 and 2.

Representative game: $g_{67} = \begin{bmatrix} 4,1 & 2,3 \\ 1,4 & 3,2 \end{bmatrix}$

Although the theoretical L$k(\alpha)$ classes do not predict all the empirical components correctly, there are important regularities, which are not predicted by other theoretical concepts: (*i*) The OD1 and OD2 games are never in the same empirical component, as is predicted by Lk(α) (unlike the Nash and strict dominance similarity classes). (*ii*) Similarly, the CO1 and CO2 games are never in the same empìrical components, as is predicted by L$k(\alpha)$ (again unlike the Nash and strict dominance similarity classes). Finally, (*iii*) All but one MP games remain in the same component. This is the only prediction, which is also predicted by NE and strict dominance similarity classes.

However, as is known from the experimental literature, the PD games and the coordination games with Pareto-rankable outcomes (games with outcome with payoffs (4,4)) and the BoS and Chicken games are at the center of attention and Lk-models do not pick-up tension with being or not a NES. At the same time, the experimental literature does not systematically study the whole variety of related games, especially asymmetric PD, stag hunt, or BoS games.

B. Empirical similarity classes: Breakaway components

We have seen how 4 of the 14 games that break away from the theoretical L$k(\alpha)$ classes join the OD2 component. We now study how four new empirical classes are formed by the 10 remaining "breakaway" games. Three classes consist of games from the theoretical L$k(\alpha)$ one-outcome class and behave as four-outcome games with actual relative frequencies between 70-50% rather than frequencies of 75-75% or higher, associated with the one-outcome class A1 (as mentioned above; see Figure 7, central area and top-right corner).

The fourth class consists of a singleton game that breaks away from the four-outcome L$k(\alpha)$ CO2 class and behaves as a one-outcome game with frequencies close to 80-80%, but with no direct neighbors in the one-outcome L$k(\alpha)$ class.

B1. Four-outcome PD-like games (4 games of which DD: 1, OD1: 3) **Features:** For both players all level-$k(\alpha)$ actions lead to the unique (pure) Nash equilibrium but do not lead to a near-equal split outcome.

Relation to L$k(\alpha)$ classes: Breakaway from the one-outcome class.

- These include the Prisoner's Dilemma game g_{57} (DD game), 2 asymmetric PD-like games g_{55}, g_{56} (OD1 games), where both players must deviate from the equilibrium predicted by L1(α) to reach a near-equal split outcome, and, finally, game g_{54} (OD1 game), where the NE is Pareto efficient, but not NES (NE=L1(α)≠NES). [26]

Representative game: $g_{57} = \begin{bmatrix} 4,1 & 2,2 \\ 3,3 & 1,4 \end{bmatrix}$

B2. Four-outcome Stag Hunt-like games (3 games all CO1) **Features:** For both players all level-$k(\alpha)$ actions lead to the same (pure) Nash equilibrium but do not lead to a near-equal split outcome; moreover, the near-equal split outcome is a Nash equilibrium.

Relation to L$k(\alpha)$ classes: Breakaway from the one-outcome class.

- This component consists of the 3 games g_{69}, g_{70}, g_{76}, where the outcome predicted by L1(α) corresponds to a Pareto-inferior Nash equilibrium. In contrast, the other Nash equilibrium (not predicted by L1(α)) is a NES and thus, Pareto-dominant. Both players must switch actions from the L1(α) ones to achieve the payoff improvement, which approximately half of the subjects do (see Figure 7). This component is closest to the midpoint frequency (50-50%) and is distinct from the PD class - where only 30% switch (as observed in the experimental literature) - despite similarities in the behavioral Lk(α) and NES structure.

Representative game: $g_{69} = \begin{bmatrix} 4,4 & 1,3 \\ 3,1 & 2,2 \end{bmatrix}$

B3. Four-outcome BoS-like games (2 games both CO1) **Features:** For both players the level-1(α) actions lead to the same (pure) Nash equilibrium that is efficient (or also near-equal split), while there is another Nash equilibrium that is near-equal split.

Relation to L$k(\alpha)$ classes: Breakaway from the one-outcome class.

- The 2 games g_{63} and g_{66} are one-outcome games based on the theoretical level-$k(\alpha)$ actions; they remain together and all four outcomes, including the two Pareto efficient NE (at least one of which is NES) are played empirically.

Representative game: $g_{66} = \begin{bmatrix} 4,2 & 2,1 \\ 1,3 & 3,4 \end{bmatrix}$

B4. One-outcome symmetric BoS-like game (One game CO2) **Features:** For both players the L1(α) action leads to a (pure) Nash equilibrium that is near-equal split but is a different

[26] Experimental studies of one-shot PD games (like g_{57}) confirm cooperation rates around 30% (see Figure 7). However, the literature does not extensively discuss which adjacent games share similar game-theoretic properties and produce similar behavior, aside from Bruns (2015) and Robinson and Goforth's (2005) theoretical discussions of such games.

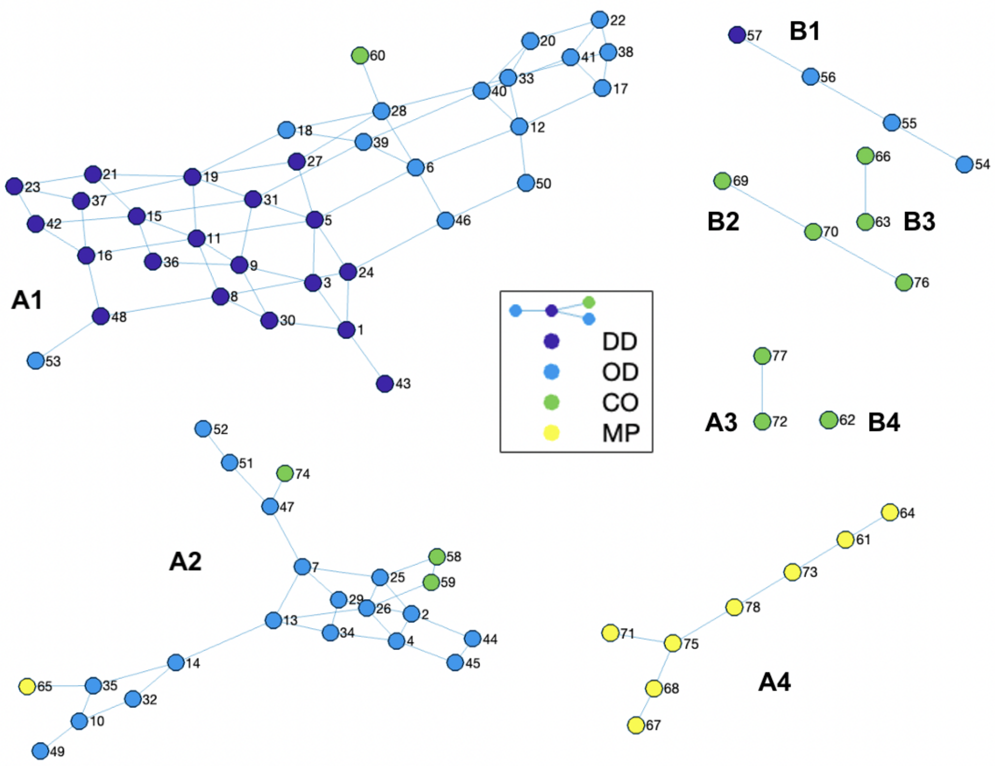


Figure 6: Graph of the empirical similarity classes in G^* from aggregate behavior of all subjects, for $\delta = 5\%$ at a 99.9% confidence level. Each node is a game in G^* and a link is drawn between a pair of games if: (i) for some representation of the pair of games in G, both games are neighbors and (ii) both games belong to the same empirical similarity class. A1-A4 are the four level-$k(\alpha)$ related components and B1-B4 are the four breakaway components.

one for the two players.

Relation to L$k(\alpha)$ classes: Breakaway from the four-outcome CO2 class.

- The game g_{62} is a four-outcome game based on the theoretical level-$k(\alpha)$ actions, yet the relative frequencies are around 80% of playing L1(α) and potential payoffs of 4 and 2 for both players, so that it behaves empirically like a one-outcome game. But note that the L1(α) actions do not coordinate on a Nash equilibrium and the game has no neighbors with the one-outcome component A1 above. Since behavior is discontinuous with respect to the other BoS games, it forms a singleton component.

Representative game: $g_{62} = \begin{bmatrix} 4,3 & 2,2 \\ 1,1 & 3,4 \end{bmatrix}$

Comparison with further key theoretical similarity classes

Lk(α) based classes. The similarity classes based on Lk(α) appear effective in organizing the empirical data, with relatively few games either breaking away from their original components to form separate components or merging with one of the main components as detailed above. Overall, the Lk(α) classes misclassify $18\%(= 14/78)$ of the games, namely, 1 DD game, 3 OD1 games, 9

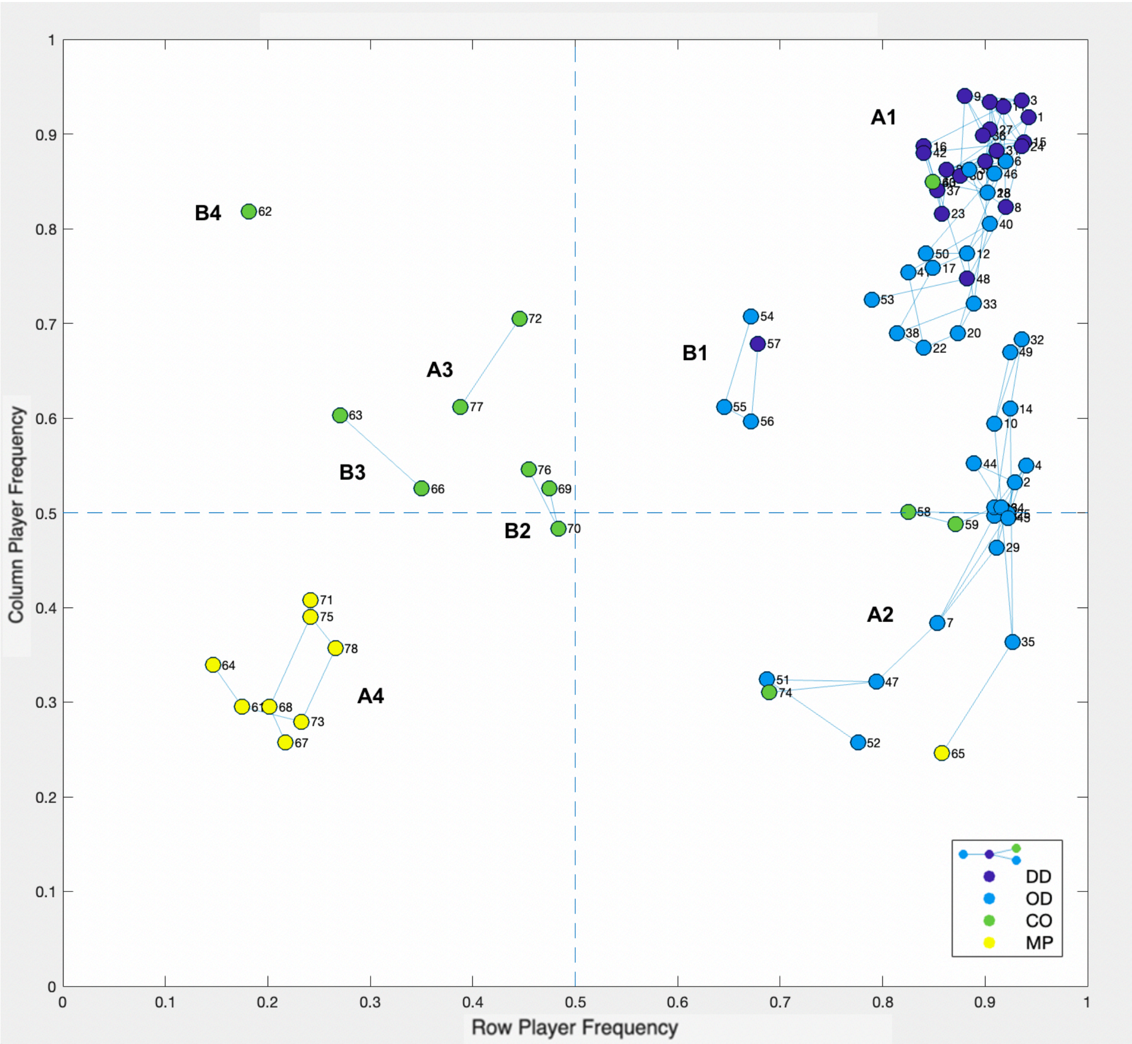


Figure 7: Empirical similarity classes with frequencies of play (renormalized for visibility). Each node is a game in G^* and a link is drawn between a pair of games if: (i) for some representation of the pair of games in G, both games are neighbors and (ii) both games belong to the same empirical similarity class.

CO games (5 CO1 and 4 CO2) and 1 MP game.[27]

Nash equilibrium and strict dominance classes. Like the Lk(α) based rules, the Nash Equilibrium and the Strictly Dominance rules successfully predict the MP class (except for game g_{65}). However, both fail to distinguish OD1 from OD2 games, for which Lk(α) provides an empirically more accurate prediction. Additionally, Lk(α) successfully distinguishes among the four (asymmetric) PD games, separating game 51 from the other three. This distinction is confirmed empirically: two asymmetric PD games, game 55 and game 56 (both OD1 games), form a component with game 57 (a PD game), separate from the one-outcome games, while game 51 belongs to the OD2 component. Overall, the Nash equilibrium classes misclassify 41%(= 32/78) of the games, namely, 1 DD game, 21 OD games, 9 CO games and 1 MP game. The strict dominance classes, on the other hand, do slightly better at classifying the OD games and misclassify 37%(= 29/78) of the games, namely, 1 DD game, 18 OD games, 9 CO games and 1 MP game.

NES and Pareto efficiency. While NES and Pareto efficiency considerations play a role in the empirical similarity classes, taken by themselves, neither of these concepts does well in explaining the empirical similarity classes, as can be seen from inspecting the theoretical classes they imply, see Figures A17 and A13 in the Appendix. They misclassify over 50% of the games.

In summary, the empirical similarity classes appear to be driven by Lk(α) rules together with features of NES, following an outcome selection logic. This raises the question: do subjects behave according to the behavioral rules identified in our analysis of similarity classes? In the remainder of this section, we address this question by demonstrating that aggregate behavior alone cannot fully rationalize these similarity classes. Instead, a closer examination of individual behavior and its alignment with certain behavioral rules provides useful insights.

6.2 Further measures of aggregate behavior

If aggregate behavior were sufficient to explain the empirical similarity classes, one might expect that examining the behavioral rules that best fit the aggregate data would suffice.

A common approach, as in Fudenberg and Liang (2019), is to assess the accuracy of each rule in predicting the *modal* play—the most frequently chosen action—in each of the 144 perspectives. However, while this approach identifies the best-performing rule for predicting modal play, it leaves unexplained residuals that may be attributable to other rules. In our analysis, the rule level-1(α) achieves the highest accuracy, correctly predicting 94.44% (136/144 perspectives) of modal play.[28] This is closely followed by level-5(α) and level-5, with statistically comparable performance (see Table 1 or Figure 8). For the eight perspectives where level-1(α) fails to predict modal play, both

[27] To compute the number of misclassified games according to a theoretical rule, we compute for every separate theoretical component or similarity class, the maximum number of games that empirically belong to that component. All other games that do not belong to that component are considered to be misclassified.

[28] Many experimental studies have found this result, e.g., Fudenberg and Liang (2019) in a meta study on 3x3 games, Selten et al. 2003 on randomly chosen 3x3 games with the strategy method, Ert et al. (2011) in simple extensive form games.

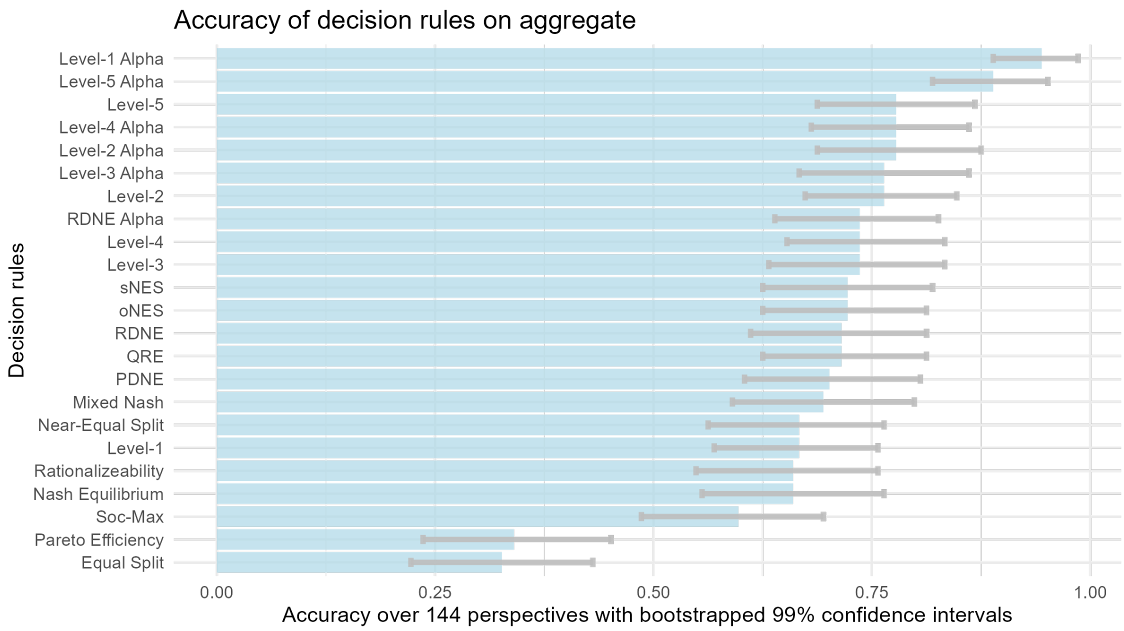


Figure 8: Accuracy of decision rules in predicting the most frequently played action.

sNES and PDNE achieve perfect accuracy, predicting 100.00% (8/8 perspectives). These findings align with the results of Fudenberg and Liang (2019) for 3×3 games, where most modal play is predicted by level-1(α), but in games where level-1(α) performs poorly, modal play often aligns with PDNE. [29]

Although level-1(α) accurately predicts modal play in over 90% of the perspectives, its similarity classes fail to align with the empirical similarity classes. Figure A6 in the Appendix illustrates that level-1(α) does not separate any games in G, as it predicts only one action per perspective. This suggests that predicting the most frequently played action may not fully capture the structure of the empirical similarity classes, which are based on whether the differences in how people play two games exceed a specified threshold, rather than solely on modal play. A more appropriate approach might involve predicting the full frequency distribution of actions and evaluating the mean-squared error (MSE) for each rule.

Under this criterion, the best-performing rule is unequivocally level-1 (see Figure 9). Other rules, such as level-1(α) and Nash equilibrium, perform similarly in terms of MSE against the data, with no statistically significant differences. Unlike level-1(α), level-1 separates games into multiple similarity classes (see Figure A9 in the Appendix). However, its partitioning of games is still less consistent with the empirical similarity classes compared to level-k(α). The adjusted Rand index (ARI) between level-1 and the empirical similarity classes is 0.3603 (99% bootstrap confidence interval: 0.2356, 0.5661; non-adjusted Rand index: 0.7409). By contrast, the ARI between level-k(α) and the empirical similarity classes is significantly higher at 0.6382 (C.I.: 0.4427, 0.7158; n.a.

Table 1: Accuracy of decision rules in predicting the most frequently played action.

Decision Rule	Games correctly predicted	99% CI lower bound	% of 144	99% CI upper bound	% of games if L1(α) incorrect
L1(α)	136	0.89	0.94	0.99	0.00
L5(α)	128	0.81	0.89	0.95	0.62
L2(α)	112	0.69	0.78	0.86	0.75
L4(α)	112	0.71	0.78	0.85	0.75
L5(α)	112	0.71	0.78	0.85	0.75
L2	110	0.69	0.76	0.83	0.75
L3(α)	110	0.70	0.76	0.83	0.62
L3	106	0.65	0.74	0.83	0.75
L4	106	0.64	0.74	0.80	0.75
oNES	104	0.63	0.72	0.81	1.00
sNES	104	0.62	0.72	0.81	1.00
RDNE	103	0.62	0.72	0.81	0.75
QRE	103	0.62	0.72	0.81	0.50
PDNE	101	0.59	0.70	0.80	1.00
NES	96	0.55	0.67	0.76	1.00
L1	96	0.56	0.67	0.74	0.00
NE	95	0.55	0.66	0.76	0.62
RAT	95	0.56	0.66	0.74	0.62
Soc-Max	86	0.49	0.60	0.69	1.00
PE	49	0.24	0.34	0.46	0.88
ES	47	0.21	0.33	0.40	0.88

RI: 0.8362).[30]

Additionally, Figure A28 in the Appendix reports both MSE and accuracy simultaneously, providing a detailed comparison of rule performance across perspectives. This figure highlights the trade-offs between modal play accuracy and distributional fit for the various rules.

Furthermore, neither level-1, level-1(α), nor level-k(α) similarity classes distinguish the most well-known games, such as Prisoner's Dilemma (PD), as behaviorally distinct from other double-dominance (DD) games.

Since no single rule or family of rules fully explains the empirical similarity classes, it is plausible that subjects adopt different rules. The empirical similarity classes may therefore reflect heterogeneity in subjects' behavior rather than a single unifying rule.

6.3 Individual behavior and behavioral rules

Given that the empirical similarity classes depend on the proportions of rule adoption among subjects, we analyze the behavior of each of the 450 participants across the 144 game perspectives. Our goal is to determine whether a subject predominantly follows a single decision rule, combines multiple rules, or plays in a manner consistent with randomness.

For each subject, we estimate their best-fitting decision rule by counting the number of choices that

[29] However, Camilo and Nagel (2023) demonstrate that NES outperforms PDNE in these 3×3 games.

[30] The adjusted Rand index (ARI) measures the similarity between two partitions, adjusting for the possibility of matching by chance.

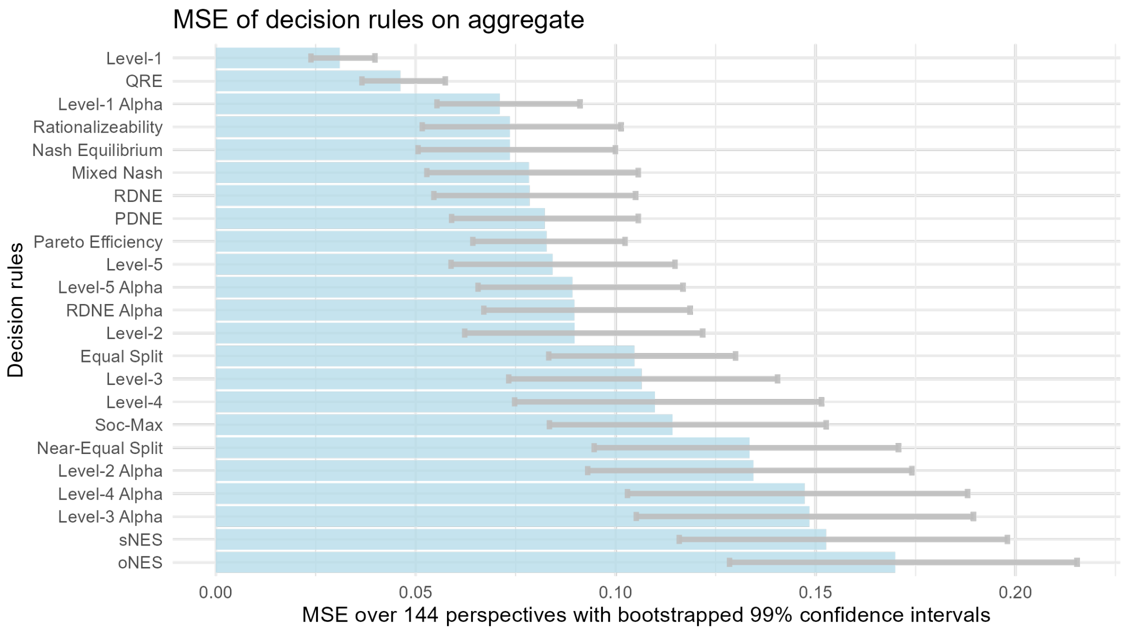


Figure 9: MSE of decision rules in predicting the frequency distribution of perspectives.

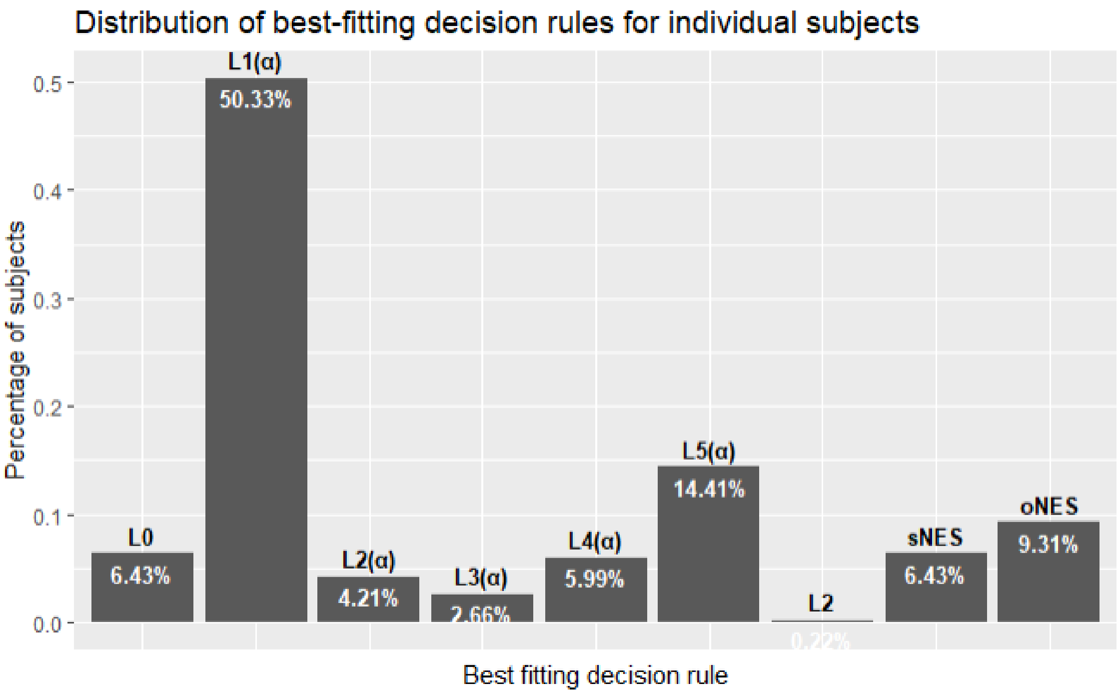


Figure 10: The distribution of all 450 subjects by best-fitting rule. Note that L5(α) is a rule identical to choosing NE action when it is unique and L1(α) otherwise

Table 2: Adjusted Rand Index of Theoretical Similarity Classes vs. Empirical Similarity Classes

Decision Rule	ARI of Rule's Classes vs. Empirical Classes	99% CI Lower Bound	99% CI Upper Bound	Figure Ref.
Data	1.00	1.00	1.00	Fig. 6
L$k(\alpha)$	0.68	0.49	0.87	Fig. 4
Lk	0.55	0.43	0.75	Fig. A8
L1	0.37	0.23	0.61	Fig. A9
NE[1]	0.36	0.20	0.55	Fig. 3
RDNE	0.33	0.19	0.51	Fig. A14
NSD	0.32	0.18	0.55	Fig. 2
L4	0.32	0.18	0.51	Fig. A11
PDNE	0.30	0.15	0.48	Fig. A15
L2[2]	0.22	0.07	0.41	Fig. A10
L2(α)[3]	0.08	0.00	0.20	Fig. A12
NES	0.03	-0.08	0.16	Fig. A13
Soc-Max	0.02	-0.08	0.20	Fig. A19
QRE	0.00	0.03	0.10	Fig. A16
L1(α)[4]	0.00	0.00	0.00	Fig. A6
ES	-0.01	-0.07	0.10	Fig. A18
PE	-0.02	-0.04	0.10	Fig. A17

[1] RAT produces the same classes. [2] L3 and L5 produce the same classes. [3] L3(α), L4(α), and L5(α) produce the same classes. [4] oNES, sNES, and Max-Max produce the same classes.

align with each rule. The rule with the highest number of correctly predicted decisions is considered the subject's best-fitting rule. A subject is classified as following the level-0 (random) rule, which serves as our benchmark, if their best-fitting rule cannot be shown to perform better than random choice. This determination is made for a subject if we fail to reject that the best-fitting rule has an accuracy greater than 50% at a 95% confidence level.

The distribution of subjects' best-fitting decision rules is summarized in Figure 10. level-1(α) emerges as the most common best-fitting rule among individual subjects (as opposed to modal play), followed by the two near-equal split (NES) rules and level-5(α). Notably, PDNE does not appear as the best-fitting rule for any subject. Subjects whose behavior is best fit by level-1(α) exhibit a high degree of consistency in following this rule. However, when they deviate, their choices tend to align with near-equal split, as illustrated in Figure 11. Interestingly, this pattern of deviations works in reverse as well, indicating that the presence of level-1(α) and near-equal split among the subject distribution is not merely reflective of two distinct "types" of players but rather a spectrum where subjects combine these two rules to varying degrees.[31]

For most subjects best fit by level-1(α) (or near-equal split), more than 90% of their choices can be explained by combining these two rules. Moreover, combining level-$k(\alpha)$ rules with near-equal split significantly improves predictive accuracy for individual behavior, a result that is statistically robust and not easily matched by other rule combinations (see Figure A26 in the Appendix for a complete version of Figure 11).

[31] Figure A27 in the Appendix shows that essentially the only other rule that plays a role as a best-fitting secondary rule at the individual subject level, beyond the mentioned Level-$k(\alpha)$ and NES rules, and for a very small number of subjects ($< 5\%$), are the Soc-Max and Nash equilibrium rules.

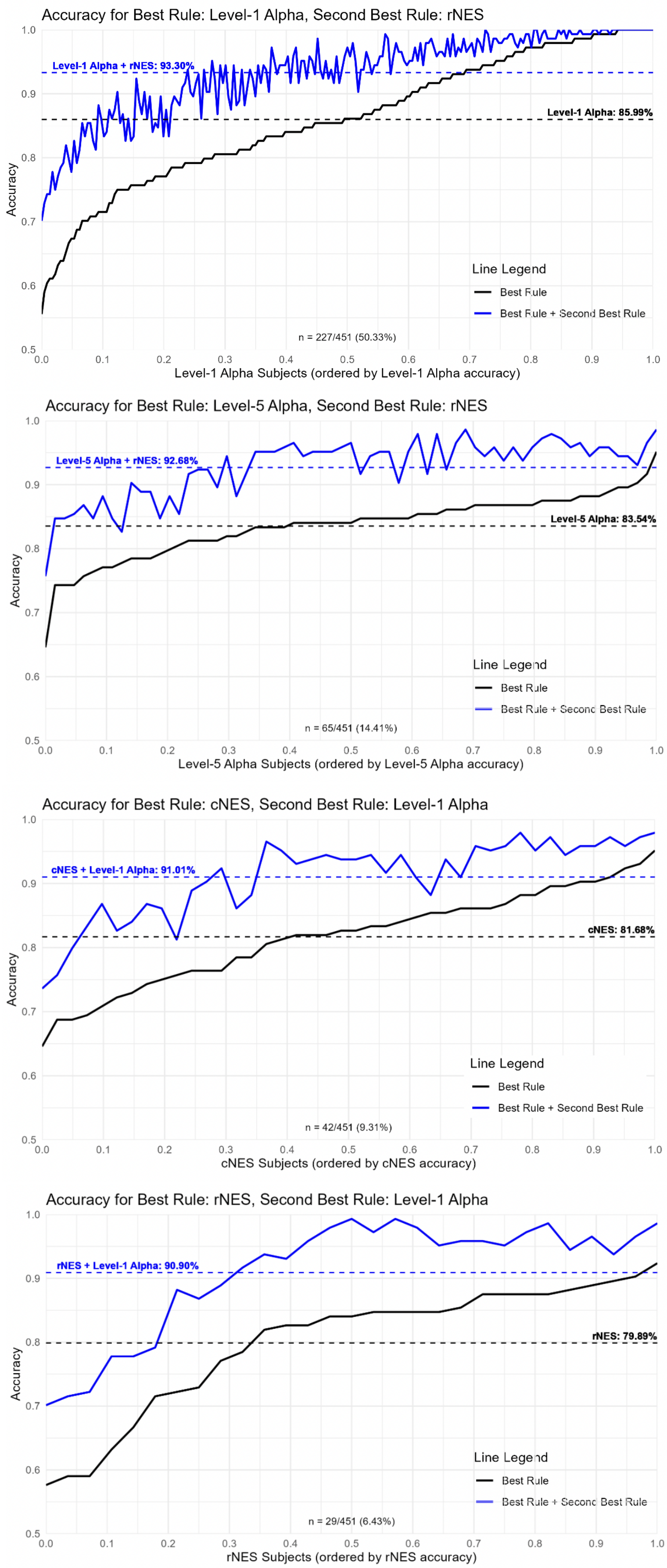


Figure 11: Accuracy of each subject's best-fitting rule (black line) plus added accuracy of a second rule, computed as the best-fitting rule over those games, where the subject's best-fitting rule predicts incorrectly (blue line). The figures report the four most popular best-fitting rules, namely, L1(α), L5(α), cNES and rNES with second rules, respectively, rNES, rNES, L1(α) and L1(α).

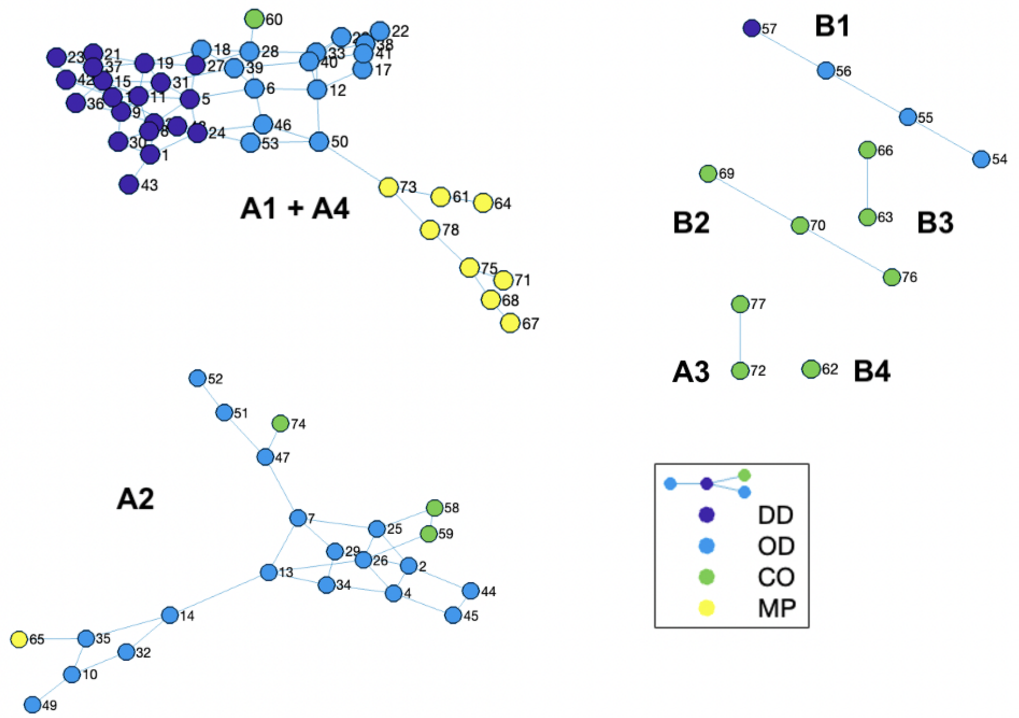


Figure 12: Graph of the empirical similarity classes in G^* from aggregate behavior of just those subjects that were identified as being level-$k(\alpha)$ types, for $k = 2, \ldots, 5$, and for $(\Delta, p) = (5\%, 99.9\%)$.

The complementarity between level-$k(\alpha)$ and near-equal split provides a framework for understanding why the empirical similarity classes further refine the similarity classes implied by level-$k(\alpha)$. For instance, the Stag Hunt class is separated from the BoS class because the near-equal split equilibrium is particularly focal—not only for the 15% of subjects best fit by near-equal split rules, but also for other subjects who frequently deviate from their primary rule to play near-equal split. Similarly, PDs are separated from the rest of the DD games because the Nash equilibrium in PDs is not Pareto-optimal, while a Pareto-optimal NES exists that is appealing to NES-types and also to some level-$k(\alpha)$ types.

6.4 Individual behavior and similarity classes: A first application

To better understand the type of games at which some of the subjects identified as level-$k(\alpha)$ switch from following a level-$k(\alpha)$ rule to following a near-equal-split rule, we compute the empirical similarity classes for all level-$k(\alpha)$ types of subjects together but also for subgroups of level-$k(\alpha)$ types. Surprisingly, as can be seen in Figure 12, the "higher" level-$k(\alpha)$ types, that is with $k = 2, .., 5$ have essentially exactly the same similarity classes as the ones obtained from all 450 subjects (see Figure 6). The difference with the similarity classes of all subjects is one extra link connecting the A1 (one-outcome) and the A4 (MP) components, which for this group of players form a single similarity class (A1+A4). Taking just the level-1(α) types leads to a unique similarity class consisting of all games (see Figure A23 in the Appendix). At the same time, taking all level-$k(\alpha)$

types together ($k = 1, \ldots, 5$) leads again to similarity classes that are also very close to the one of all subjects (see Figure A24 in Appendix) but not as close as the one for the higher level-$k(\alpha)$ types. The appendix also contains the similarity classes for subjects that are classified as near-equal-split types (see Figure A25 Appendix). In particular, this suggests that the higher level-$k(\alpha)$ types, while following the level-$k(\alpha)$ rules on a large subset of games, appear to deviate from those rules precisely on the prisoner's dilemma, stag-hunt and battle-of-the sexes like games (B1, B2, B3, B4 components). Moreover, because the other player types (level-$1(\alpha)$ and near-equal-split) have different similarity classes (though not totally different ones), it seems that it is the choices of these higher level-$k(\alpha)$ types that make the crucial difference in determining the similarity classes for the overall population of subjects.

7 Discussion and directions for future research

This study has presented a framework for grouping games based on the continuity of behavior, taking this as a basis to understand human behavior on a comprehensive domain of 2×2 games. We first studied how different concepts from game theory and behavioral game theory imply similarity classes based on the continuity of recommended or predicted behavior. We then studied empirical similarity classes and discussed how they differ from the theoretical ones. However, perhaps not too surprisingly, the level-$k(\alpha)$ rules provide the closest fit. Some exceptions occur when the level-$k(\alpha)$ actions correspond to Nash equilibrium actions that are not Pareto efficient or not fair. In such cases the NES rules appear to be relevant. Confronting subjects with all 144 perspectives revealed that most subjects' behavior depends on a mix of level-$k(\alpha)$ and NES rules, albeit to varying degrees. Such diversity at the individual subject level cannot be easily uncovered when studying small domains of games. Inconsistencies in level-k rules across different games from the literature (Georganas, Healy and Weber, 2015) suggest a need to rethink this common approach. Nonetheless, several extensions and refinements remain to be explored that can strengthen and expand the current analysis. In the following, we outline such areas for further investigation.

7.1 Insights from similarity classes

Generalizability of canonical games. The similarity classes computed in this study based on the continuity of behavior allow indirectly to quantify the extent to which behavior generalizes and remains predictable across games. Interestingly, canonical games, such as the prisoner's dilemma, stag hunt, battle of the sexes, and matching pennies games that provide key insights into strategic interaction are all singled out as separate classes by the empirical classification. At the same time, they constitute relatively small classes within our domain.

For example, the symmetric battle of the sexes game belongs to a similarity class comprising only 2/78 games (2.6%), while the matching pennies class (excluding g_{65}) spans 8/78 games (10.3%). By contrast, the one-outcome class encompasses 36/78 games (46.2%), and the two-outcome class 22/78 games (28.2%). This shows that insights derived from canonical games such as battle of the sexes or matching pennies come from classes that represent relatively special cases. Behavior may be predictable within these small classes, but the lessons do not necessarily generalize to the much larger classes in the 2×2 domain. These larger classes deserve closer attention—also from a mechanism design perspective—because they are characterized by the desirable property that play concentrates reliably on one or two outcomes.

Our contribution extends this discussion by systematically analyzing *all* possible games in this context, in particular, many asymmetric games nearby canonical games. This allows us to identify which games produce empirically similar behavior and which do not, as well as to clarify how canonical games are grouped into the various classes. For instance, coordination games can be split depending on whether level-1(α) reasoning coincides or not with a Nash equilibrium, bringing them into different empirical classes. At the same time, our analysis brings to light several understudied games—such as asymmetric prisoner's dilemma type and asymmetric battle of the sexes type games—that deserve further investigation both experimentally and theoretically.

Brandenburger and Nalebuff (1995) argue that strategic success often comes from changing the game rather than merely playing it. Building on that insight, we propose focusing on the neighboring games of hard-to-predict canonical games: small, targeted payoff changes that move a game into a different class with more desirable equilibrium properties or simpler level-k rules with favorable outcomes. Suppose we can identify nearby games whose structure yields better (welfare-maximizing, more stable, or more cooperative) outcomes. In that case, modest manipulations of payoffs — incentives, transfers, information shifts, or institutional tweaks — can transform the original game into one with reliably better outcomes. This approach treats game design as a local search in payoff space: find a nearby game with favorable (empirical) outcomes, not necessarily equilibrium, and implement minimal interventions that shift incentives toward that neighbor.

Comparing solution concepts. Similarity classes also provide a framework for assessing the generalizability of solution concepts. By analyzing how well different behavioral rules predict choices within each class, we can evaluate the relative strengths and weaknesses of these concepts. Preliminary findings in the Appendix indicate that level-1 and level-1(α) perform consistently well across most similarity classes, likely due to their alignment with the structure of this domain (see Figures A29 to A32). Future work could involve analyzing rule performance within and across similarity classes more systematically. For instance, comparing the predictive power of near-equal split and level-1 rules in classes containing games that clearly separate them may yield deeper insights into how these rules complement each other.

7.2 Robustness and generalization

Payoff structures. The robustness of our results to different payoff structures remains an open question. Our analysis relies on payoffs $\{1, 2, 3, 4\}$, which are treated as cardinal values by subjects. Exploring alternative payoff sets would help clarify whether behavioral rules and similarity classes

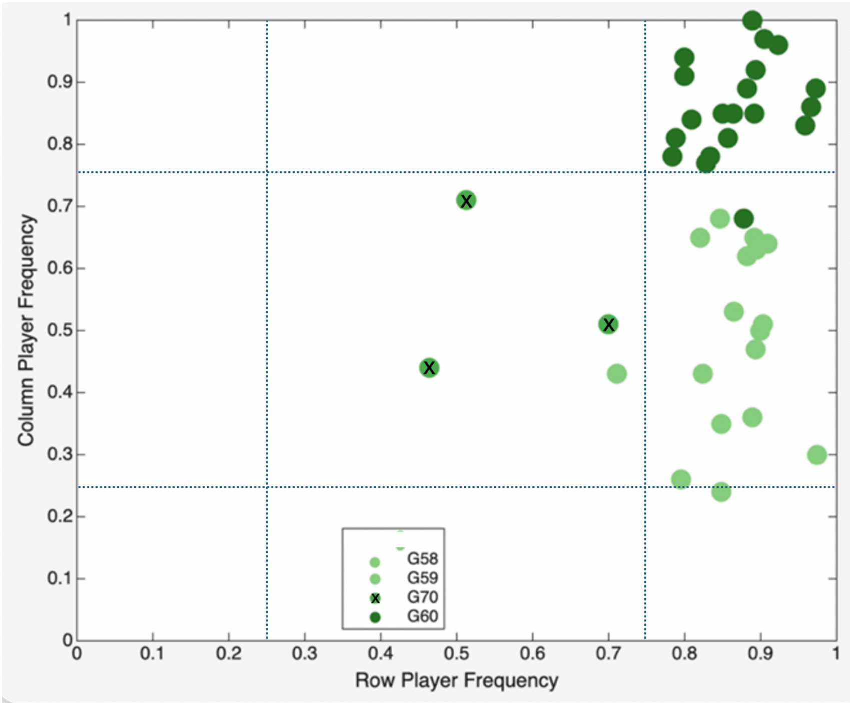


Figure 13: Frequencies of play (renormalized for visibility) in games taken from Zhu et al. (2025) with payoffs in [0, 50] and payoff structures corresponding to the ones of games g_{58}, g_{59}, g_{60} and g_{70} respectively.

remain consistent across different payoff scales. In a companion paper, we are currently analyzing the data from Zhu et al. (2025), a rich data set on 2×2 games with payoffs systematically drawn from the interval [0, 50] based on the Robinson and Goforth (2005) classification, without matching pennies-type games.

To briefly illustrate the potential of our approach, we use the similarity classes obtained in Section 6.1 to predict behavior in the more general domain of 2×2 games studied by Zhu et al. (2025) for the four coordination games g_{58}, g_{59}, g_{60} and g_{70} discussed in the Introduction. for this, we take all games from the Zhu et al. (2025) data set that have same order of payoffs as the games $g_{58}, g_{59}, g_{60}, g_{70}$, and the level-$k$ (and level-$k(\alpha)$ for any $\alpha < 1$) predictions coincide with the ones of those games. In particular, we expect to see the games with the structure of g_{60} in the one-outcome quadrant $[0.75, 1] \times [0.75, 1]$ (as A1 of Figure 7), the games with the structure of the games g_{58}, g_{59} in the two-outcome quadrant $[0.75, 1] \times [0.25, 0.75]$ (as A2 of Figure 7), and games with the structure of g_{70} in the four-outcome quadrant $[0.25, 0.75] \times [0.25, 0.75]$ (as B2 of Figure 7). Of the 40 games from Zhu et al. (2025) that have payoffs that correspond to the ones respectively of $g_{58}, g_{59}, g_{60}, g_{70}$ mentioned above, 38 have frequencies that are in the expected quadrant consistent with their respective empirical similarity classes: 15/16 for games like g_{58} and g_{59} in A2, 19/20 for games like g_{60} in A1, and 3/3 for games like g_{70} in B2. See Figure 13.

Expanding the behavioral rule set While our analysis includes a broad set of behavioral rules, there are omissions that warrant further exploration. For example, *quantal response equi-*

librium (QRE, McKelvey and Palfrey, 1995), which captures stochastic decision-making under uncertainty, may better explain behavior in certain games, particularly with mixed strategy equilibria, like matching pennies type games. However, our first analysis across all games suggests that the original QRE model does not explain the emprical similarity classes well (Figure A16 in the Appendix). Incorporating extended QRE and other models, including hybrid rules combining multiple concepts, such as fairness and probabilistic choice, may account for the observed deviations and heterogeneity of behavior at the subject level. Developing such extensions requires designing experiments to test these models and refining methodologies to accommodate more complex rule combinations.

Sample size and fatigue. Subjects in our study played 144 perspectives, raising concerns about fatigue or noise. Results reported in the Appendix show that the distribution of behavior remains stable across the three blocks of games (1–48, 49–95, 96–144), suggesting minimal order effects (see Figures A33 to A35). Further analyses, such as regression models testing the influence of sequence assignment on choices and rule fit, would help formalize these robustness checks.

7.3 Applicability to broader game domains

Beyond 2×2 games. The methods developed in this paper can, in principle, be applied to other game domains, such as 3×3 games or parametric spaces of games, such as Cournot or beauty contest games. However, computational challenges emerge as the size of the game space increases. For instance, while the 2×2 domain contains 576 games, the 3×3 domain includes $(9!)^2 \approx 10^{12}$ games. Algorithmic advances would be required to efficiently compute neighborhood structures and similarity classes in these larger domains. The paper by Fudenberg and Liang (2019) provide such an approach. In a paper by Camilo and Nagel (2025) we show that the concept of NES provides a better fit to one of the data sets than the newly derived Pareto-Dominant Nash equilibrium (PDNE).

Parametric game spaces. In parametric spaces, such as generalized beauty contests or linear best-response games, defining a topology based on parameter distances offers a promising strategy for extending the framework. This approach would allow researchers to analyze how strategic behavior varies with changes in parameters and identify structural regularities. Subjects may be asked to play all games from a discretized subset of the parametric space. The theoretical tools introduced here, including the computation of similarity classes, seem well-suited for such applications.

7.4 Optimal play and complexity

Identifying optimal strategies. The empirical frequency distributions collected in this study provide a foundation for identifying optimal strategies. Preliminary results suggest that level-$1(\alpha)$ aligns closely with optimal play in dominance-solvable games, while four-outcome games like matching pennies require more sophisticated strategies. The findings indicate that the domain of 2×2 games varies widely in terms of strategic complexity.

Response time and strategic complexity. An inverse relationship between response times and optimal play suggests that 2×2 games are perceived as low-complexity problems by subjects. In the Appendix, we find that this relationship holds in our data (see Figure A36). This finding aligns with the literature on speed, accuracy, and complexity (Goncalves, 2024). Analyzing how response times vary across similarity classes could provide further insights into the cognitive processes underlying strategic decision-making.

8 Conclusion

This paper takes two new steps. First, it examines human behavior across a comprehensive domain of one-shot 2×2 games with each subject deciding on all games of the domain, making it one of the largest such experiments on one-shot games. Second, it uses a geometric approach to group games into theoretical and empirical similarity classes. These similarity classes capture both the predictions of behavioral or game-theoretic solution concepts as well as observed behavior from participants on the Prolific platform. We find that empirical similarity classes, derived from aggregate data, align closely with theoretical similarity classes based on level-$k(\alpha)$ rules, with near-equal split (NES) rules playing a complementary role, particularly in cases where NES outcomes differ from level-$k(\alpha)$ or Nash equilibrium outcomes. The capacity of level-$k(\alpha)$ rules to explain individual behavior is demonstrated across a broader range of games than previously documented, further supporting their applicability across diverse strategic contexts. Given the large number of games observed per subject, behavior is identified that follows multiple behavioral rules, instead of the usual one-rule plus noise studied in most of the literature.

The aggregate behavior observed on our domain can be viewed as providing a reference library of how human subjects play 2×2 games with payoffs from $\{1, 2, 3, 4\}$ without replacement. This "behavioral atlas" provides a starting point for researchers and practitioners interested in predicting, comparing, or modeling human decisions in such environments. By systematically mapping the empirical regularities across the various classes of games, the study offers both a benchmark for future theoretical work and a foundation for extending behavioral models to broader domains. It also opens the door to game design applications. Following Brandenburger and Nalebuff's (1995) insight that strategic success often comes from changing the game rather than merely playing it, our atlas enables the identification of "neighboring games" of hard-to-predict canonical ones. Small adjustments to payoffs, incentives (or institutions) could shift such games into classes with more desirable properties, offering a systematic way to design "better" games through an informed local search in payoff space.

Our analysis also clarifies the predictive roles of different solution concepts on a broad domain. The shortcomings of the Nash equilibrium and strict dominance compared to level-$k(\alpha)$ are apparent in their classifications especially of dominance solvable games. Whereas strict dominance separates one-sided from two-sided dominance games, Nash equilibrium does not separate them at all. At the same time, the empirical similarity classes suggest that the more relevant separation is the one between games solvable in one or two steps of level-$k(\alpha)$ reasoning, that is, DD and OD1 games on one side and OD2 games on the other. Matching pennies games, by contrast, are correctly classified by all concepts into one separate empirical similarity class, with one game exception. As discussed in the experimental literature, coordination games remain among the most challenging to predict. Level-$k(\alpha)$ distinguishes two classes of coordination games, depending on whether the Nash equilibrium is reached in one step or in more than one step of level-$k(\alpha)$ reasoning. Strict dominance and Nash equilibrium subsume them all into one separate class. Efficiency and fairness considerations appear to play a role in coordination that is not picked up by either strict dominance or Nash equilibrium, but also not by level-$k(\alpha)$ rules.

Looking ahead, future research should test the centrality of level-$k(\alpha)$ and NES rules in richer game contexts, investigate the stability and applicability of empirical similarity classes beyond 2×2 games, and explore more in detail the multiplicity of rules of criteria adopted by individual subjects. Identifying "special games" within classes may also clarify why certain structures or actions become focal. Together, these directions promise to refine our models of strategic reasoning and more broadly improve predictions of human decision-making.

References

[1] **Aliprantis, Charalambos D. and Kim C. Border.** 2006. *Infinite Dimensional Analysis: A Hitchhiker's Guide.* Berlin, Heidelberg: Springer Berlin Heidelberg.

[2] **Alon, Noga, Rudov, Kirill and Leeat Yariv.** 2025. "Dominance Solvability in Random Games" *arXiv preprint arXiv* 2105.10743.

[3] **Bacharach, Michael.** 2006.*Beyond Individual Choice: Teams and Frames in Game Theory.* Natalie Gold and Robert Sugden (eds). Princeton: Princeton University Press.

[4] **Biggar, Oliver, and Iman Shames.** 2023. "The graph structure of two-player games." *Scientific Reports*, 13(1): 1833.

[5] **Blume, Lawrence E. and William R. Zame.** 1994. "The Algebraic Geometry of Perfect and Sequential Equilibria." *Econometrica.* 62(4): 783-794.

[6] **Bolton, Gary E. and Alex Ockenfels.** 2000. "ERC: A Theory of Equity, Reciprocity, and Competition." *American Economic Review.* 91(1): 166-193.

[7] **Brandenburger, Adam M., and Barry J. Nalebuff.** 1995. "The right game: Use game theory to shape strategy." *Harvard Business Review.* 76: 57-71.

[8] **Bruns, Bryan.** 2015. "Names for Games: Locating 2×2 Games." *Games.* 6(4), 495-520.

[9] **Camerer, Colin and Teck Hua Ho.** 2003. "Experience-weighted Attraction Learning in Normal-Form Games." *Econometrica.* 67(4): 827-874.

[10] **Camerer, Colin, Teck Hua Hong, and Juin-Kuan Chong.** 2004. "A Cognitive Hierarchy Model of Games." *Quarterly Journal of Economics.* 19(3): 861-898.

[11] **Camilo Moore, Amil and Nagel, Rosemarie.** 2025. "Not uncovered by the machine learning algorithm: equal split as a (best) predictor in initial play?," Universitat Pompeu Fabra, mimeo.

[12] **Carlsson, H., and E. Van Damme.** 1993. "Global games and equilibrium selection." *Econometrica.* 61(5): 989–1018.

[13] **Chen, Daniel L., Schonger, M., and Chris Wickens.** 2016. "oTree — An open source platform for laboratory, online, and field experiments." *Journal of Behavioral and Experimental Finance.* 9: 88-97.

[14] **Costa-Gomes, Miguel, Vincent P. Crawford, and Bruno Broseta.** 2001. "Cognition and Behavior in Normal-Form Games: An Experimental Study." *Econometrica.* 69(5): 1193-1235.

[15] **Crawford, V.P., Costa-Gomes, M.A., and N. Iriberri.** 2013. Structural models of nonequilibrium strategic thinking: theory, evidence, and applications. *Journal of Economic Literature.* 51(1): 5–62

[16] **Erev, Ido and Alvin E. Roth.** 1995. "Predicting How People Play Games: Reinforcement Learning in Experimental Games with Unique, Mixed Strategy Equilibria." *American Economic Review.* 88(4): 848-881.

[17] **Ert, Eyal, Ido Erev, and Alvin E. Roth** 2011. "A Choice Prediction Competition for Social Preferences in Simple Extensive Form Games: An Introduction." *Games.* 2: 257-276.

[18] **Evers, Ellen RK and Imas, Alex and Kang, Christy** 2021. "On the role of similarity in mental accounting and hedonic editing.," *Psychological Review* 29(4) : 777-789., 193: 105215.

[19] **Falk, Armin, Anke Becker, Thomas Dohmen, Benjamin Enke, David Huffman, and Uwe Sunde.** 2018. "Global Evidence on Economic Preferences." *Quarterly Journal of Economics.* 133(4): 1645-1692.

[20] **Fehr E., and K.M. Schmidt.** 1999. A Theory of Fairness, Competition, and Cooperation. *The Quarterly Journal of Economics*, 114(3): 817-868.

[21] **Fudenberg, Drew, and Annie Liang.** 2019. "Predicting and Understanding Initial Play." *American Economic Review*, 109(12): 4112-4141.

[22] **Gale, David and Nikaidô, Hukukane.** 1953. "The Determination of Subjective Probability in Experimental Situations," *Econometrica*, 21(1): 52-63.

[23] **Germano, Fabrizio.** 2006. "On some geometry and equivalence classes of normal form games." *International Journal of Game Theory.* 34(4): 561-581.

[24] **Gilboa, Itzhak, and David Schmeidler** 1995. "Case-based decision theory." *The Quarterly Journal of Economics* 110(33): 605-639.

[25] **Goncalves, Duarte.** 2024. "Speed, Accuracy and Complexity." *Working Paper.*

[26] **Grimm, Veronika and Mengel, Friederike.** "An experiment on learning in a multiple games environment," *Journal of Economic Theory*, 147(6):2220-2259..

[27] **Guida, G. and Devetag, G.** 2013. "Feature-Based Choice and Similarity Perception in Normal-Form Games: An Experimental Study," *Games*, 4(4):776-794.

[28] **Güth, Werner and Kocher, Martin G.** 2014. "Ultimatum Bargaining Experiments: A Review," in *Handbook of Experimental Economics, Volume 2*, Elsevier, pp. 247–320.

[29] **Harsanyi, John C., and Reinhard Selten.** 1988. *A General Theory of Equilibrium Selection in Games.* MIT Press.

[30] **Jehiel, Philippe.** 2005. "Analogy-based expectation equilibrium," *Journal of Economic Theory* 123: 81–104.

[31] **Mauersberger, Felix and Nagel, Rosemarie** 2018. "Levels of reasoning in Keynesian Beauty Contests: A Generative Framework.," in *Handbook of Computational Economics.* Elsevier. 4:541-634.

[32] **McKelvey, Richard D. and Thomas R. Palfrey.** 1995. "Quantal Response Equilibria for Normal Form Games." *Games and Economic Behavior.* 10(1): 6-38.

[33] **Mullainathan, Sendhil and Schwartzstein, Joshua and Shleifer, Andrei.** 2008. "Coarse Thinking and Development Economics," *Quarterly Journal of Economics*, 123(1): 263–281.

[34] **Nagel, Rosemarie.** 1995. "Unraveling in Guessing Games: An Experimental Study." *American Economic Review.* 85(5): 1313-1326.

[35] **Nash, John F.** 1950. "Equilibrium Points in N-Person Games," *Proceedings of the National Academy of Sciences*, 36(1): 48-49.

[36] **Nash, John.** 1951. "Non-Cooperative Games, *Annals of Mathematics*, 54: 286-295.

[37] **Omidshafiei, Shayegan, et al.** 2020. "Navigating the landscape of multiplayer games." *Nature communications*, 11(1): 5603.

[38] **Luce, Robert Duncan and Raiffa, Howard.** 1957. *Games and Decisions: Introduction and Critical Survey,* Wiley.

[39] **Roads, Brett. D. and Bradley C. Love** 2024. "Modeling similarity and psychological space." *Annual Review of Psychology.* 75(1): 215-240.

[40] **Rapoport, Anatol, Melvin J. Guyer, and David G. Gordon.** 1976. *The* 2×2 *Game.* Ann Arbor: University of Michigan Press.

[41] **Robinson, David and David Goforth.** 2005. *The Topology of* 2×2 *Games: A New Periodic Table.* Routledge.

[42] **Rubinstein, Ariel** 1988. "Similarity and decision-making under risk (Is there a utility theory resolution to the Allais paradox?)." *Journal of Economic Theory* 46(1): 145-153.

[43] **Schanuel, Stephen, Leo K. Simon, and William R. Zame.** 1991. "The Algebraic Geometry of Games and the Tracing Procedure." *Game Equilibrium Models II.*

[44] **Schuirmann, Donald J.** 1987. "A Comparison of the Two One-Sided Tests Procedure and the Power Approach for assessing the equivalence of average bioavailability." *Journal of Pharmacokinetics and Biopharmaceutics.* 15(6): 657-680.

[45] **Selten, R and Abbink, K and Buchta, J and Sadrieh, A** 2003. "How to Play 3x3 Games: A Strategy Method Experiment." *Games and Economic Behavior*, 45(1): 19–37.

[46] **Shepard, Roger N.** 1987. "Toward a universal law of generalization for psychological science." *Science.* 237(4820): 1317-1323.

[47] **Shubik, Martin.** 2012. "What is a Solution to a Matrix Game," Cowles Foundation Discussion Paper No. 1866.

[48] **Sugden, Robert.** 1993. ¨Thinking as a Team: Toward an Explanation of Nonselfish Behavior.¨ *Social Philosophy and Policy.* 10: 69-89.

[49] **Tversky, Amos.** 1977. "Features of similarity." *Psychological Review.* 84(4): 327.

[50] **Vives, Xavier.** 2005. *Oligopoly, Competition and Efficiency*, Oxford University Press.

[51] **Von Neumann, John.** 1928. "Zur Theorie der Gesellschaftsspiele," *Mathematische Annalen*, 100(1): 295-320.

[52] **Von Neumann, John and Oskar Morgenstern.** 1944. *Theory of Games and Economic Behavior.* Princeton University Press.

[53] **Wright, Kevin and Kevin Leyton-Brown.** 2014. "Level-0 meta-models for predicting human behavior in games," *Journal of Artificial Intelligence Research 64,* 357–83.

[54] **Zhu, Jian-Qiao, Peterson, Joshua C., Enke, Benjamin, and Thomas L. Griffiths** 2025. "Capturing the complexity of human strategic decision-making with machine learning." *Nature Human Behaviour*, 9: 2114–2120.

Appendix

A Proofs

Proof of Theorem 1. Suppose $g \sim_R g'$. Then there exists a path from g to $\psi g'$ for some ψ such that the rule is constant. Therefore, $\psi g'$ is in the same component of $(G, E(A_\Psi))$ as g'. Since $E(A_\Psi) \subset E(A_\Psi + (1 - A_\Psi) \odot A_N \odot A_R)$, this implies that there is a path from g to g' in $(G, E(A_\Psi + (1 - A_\Psi) \odot A_N \odot A_R))$. Thus, g and g' are in the same component of $(G, E(A_\Psi + (1 - A_\Psi) \odot A_N \odot A_R)$.

Suppose now that g and g' are in the same component of $(G, E(A_\Psi + (1 - A_\Psi) \odot A_N \odot A_R))$. This implies the existence of a path (g_t) such that $g_1 = g$ and $g_n = g'$ through $(G, E(A_\Psi + (1 - A_\Psi) \odot A_N \odot A_R))$. Without loss of generality, we can rewrite (g_t) as

$$(g_t) = (g_{1,1}, \dots, g_{1,\ell_1}, g_{2,1}, \dots, g_{2,\ell_2}, \dots, g_{k,1}, \dots, g_{k,\ell_k}),$$

where k is the number of times (g_t) passes through non-equivalent games and $g_{i,j} = \psi g_{i,h}$ for any $j \neq h$ for some $\psi \in \Psi$. This rewriting splits the sequence into distinct subsequences of games that are equivalent through some symmetry operation. Let ψ_i be such that $g_{i,\ell_i} = \psi_i g_{i,1}$. Consider now the following sequence:

$$(g_t^*) = \left(\Big(\prod_{s=0}^{k-1} \psi_{k-s}\Big)^{-1} \Big(\prod_{s=0}^{k-t} \psi_{k-s}\Big) g_{t,1}\right)_{t=1}^{k} = \left(\Big(\prod_{s'=1}^{t-1} \psi_{s'}\Big)^{-1} g_{t,1}\right)_{t=1}^{k} = \left(\tilde{\psi}_t g_{t,1}\right)_{t=1}^{k},$$

where we define $\tilde{\psi}_t \equiv \left(\prod_{s'=1}^{t-1} \psi_{s'}\right)^{-1}$.

We now prove that (g_t^*) is a path from g to $\psi g'$ for some $\psi \in \Psi$ such that for any consecutive games along the path, we have both $g_t^* \in N(g_{t+1}^*)$ and $R(g_t^*) = R(g_{t+1}^*)$, so that $g_t^* \sim_R g_{t+1}^*$, and so, by construction, also $g \sim_R g'$.

To see $g_1^* = g$, notice that $g_1^* = g_{1,1} = g_1 = g$. To see that $g_k^* = \psi g'$ for some $\psi \in \Psi$, notice that

$$g_k^* = \Big(\prod_{s'=1}^{k-1} \psi_{s'}\Big)^{-1} g_{k,1} = \Big(\prod_{s'=1}^{k} \psi_{s'}\Big)^{-1} \psi_k g_{k,1} = \Big(\prod_{s'=1}^{k} \psi_{s'}\Big)^{-1} g_{k,\ell_k} = \Big(\prod_{s'=1}^{k} \psi_{s'}\Big)^{-1} g_n = \psi g',$$

for $\psi = \left(\prod_{s'=1}^{k} \psi_{s'}\right)^{-1}$.

Finally, to show that both $g_t^* \in N(g_{t+1}^*)$ and $R(g_t^*) = R(g_{t+1}^*)$ hold, notice that if $g_t^* \in N(g_{t+1}^*)$, then $\psi g_t^* \in N(\psi g_{t+1}^*)$ for any $\psi \in \Psi$, and that by relabeling invariance, if $R(g_t^*) = R(g_{t+1}^*)$, then $R(\psi g_t^*) = R(\psi g_{t+1}^*)$ for any $\psi \in \Psi$. Therefore, comparing these two conditions between g_t^* and g_{t+1}^* is the same as comparing them between ψg_t^* and ψg_{t+1}^* for any $\psi \in \Psi$. And since, by construction, $g_t^* = \tilde{\psi}_t g_{t,1}$, it is also the same as comparing the conditins between $\psi_t g_{t,1}$ and $g_{t+1,1}$. Now, because $\psi_t g_{t,1} = g_{t,\ell_t}$, $(g_{t,\ell_t}, g_{t+1,1}) \in (g_t)$, and g_{t,ℓ_t} is not equivalent by symmetry operations to $g_{t+1,1}$ (by construction), it follows that $g_{t,\ell_t} \in N(g_{t+1,1})$ and $R(g_{t,\ell_t}) = R(g_{t+1,1})$ because (g_t) is a path in $(G, E(A_\Psi + (1 - A_\Psi) \odot A_N \odot A_R))$. Therefore, $g_t^* \in N(g_{t+1}^*)$ and $R(g_t^*) = R(g_{t+1}^*)$.

Since we've shown that (g_t^*) is a path from g to $\psi g'$, for some $\psi \in \Psi$, such that, for any consecutive games g_t^* and g_{t+1}^*, along the path, $g_t^* \in N(g_{t+1}^*)$ and $R(g_t^*) = R(g_{t+1}^*)$ both hold, this proves that $g \sim_R g'$. $\square$

B Additional figures

B.1 Additional graphs on the geometry of games

Equivalence of games. Figure A1 shows how the 576 payoff matrices in G reduce to the 78 games in G^* using the information from the set of symmetry operations Ψ on G and how they relate the games to one another using graph theoretic methods.

Visualizing the Robinson-Goforth topology. Figure A2 shows how the 576 payoff matrices relate to one another geometrically, representing the discrete analogue for the games in G to seeing the payoff matrices in $\mathbb{R}^8$ using the Euclidean norm.

Simplifying the set of games. Figure A3 illustrates two important insights: first, how the 78 games relate to one another using the information from the Robinson-Goforth topology, and second, how the geometry of games combined with solution concepts can generate distinct classes of games (i.e., similarity classes). For many solution concepts, it will be clear that the MP games will be separated from the CO games because there's no way to connect them to one another without passing through the arguably simpler DD and OD games (see Figure A4 for emphasis), generating discontinuities for many correspondences (e.g., Nash equilibrium, rationalizability, and level-k, to name a few).

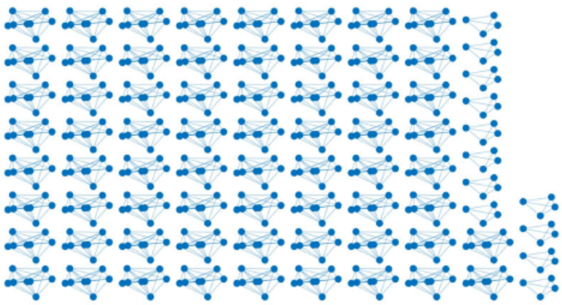


Figure A1: The graph on G, where edges represent equivalence between two games. Each of the 78 components consists of 4 or 8 equivalent games in G, depending on whether the component is one of symmetric or asymmetric games. Formally, this graph is denoted as $(G, E(A_\Psi))$.

B.2 Additional theoretical similarity classes

Level-1(α) (and other) similarity classes. Figure A6 is identical to seeing all the games in G^* in the Robinson-Goforth topology because it doesn't distinguish any similarity classes. This turns out to be the same for oNES, sNES, and Max-Max.

Level-k(α) similarity classes. Figure A5 shows the similarity classes for the set of Level-k(α) rules as shown in Figure 4, but colored to distinguish OD1 from OD2 and CO1 from CO2. The separations made by the geometry of games is further emphasized in Figure A7.

Near-Equal Split similarity classes. When considered on their own, both rules based on self-favoring or other-favoring Near-Equal Split (sNES, oNES), since they always select a unique

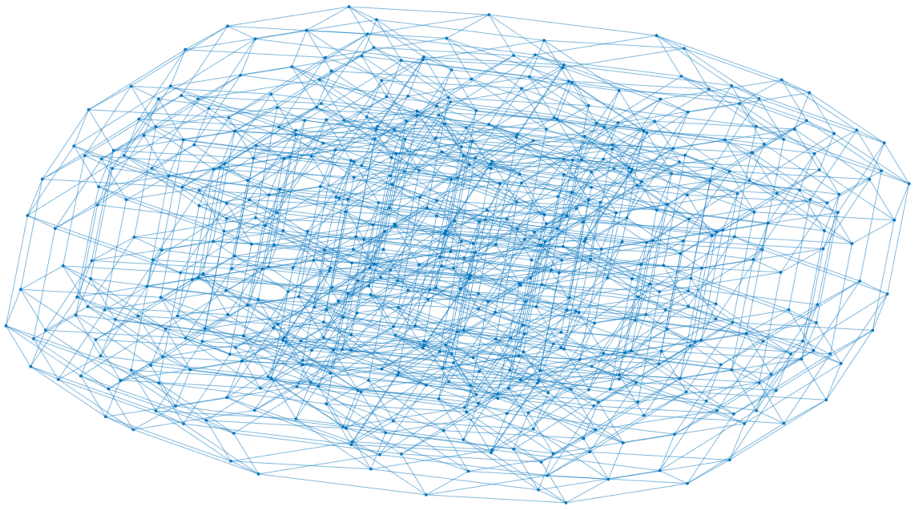


Figure A2: The graph on G, where edges represent whether two games are neighbors in the Robinson-Goforth topology. Formally, this graph is denoted as $(G, E(A_N))$.

outcome, always yield a unique similarity class, comprising all games in G (and similarly for G^*). However, taking the two rules together leads to nine similarity classes; see Figure A13. The same nine classes obtain if one takes the Near-Equal Split (NES) rule alone.

What are the distinctions between the different classes? The largest class (dark blue nodes, 65 nodes) contains all games with a unique NES which is selected by both Near-Equal-Split players. In all other games, there are two NES cells which either are in the same row, column or (off)-diagonal. The second largest similarity class (lighter blue nodes, 4 nodes) contains OD and DD games with the NES rule spanning Two-outcomes (with 4-3 or 3-4 payoffs) in the same row. The next largest class (olive green nodes, 3 nodes) has spans Two-outcomes with the 4-3 and 3-4 cells on the main diagonal and thus two actions for both players. The remaining games have two 2-3 and 3-2 or 4-2 and 2-4 outcomes in different constellations. ■

Risk dominant Nash equilibrium and Pareto dominant Nash equilibrium similarity classes. Figure A14 and Figure A15 show the similarity classes for Risk dominant Nash equilibrium and Pareto dominant Nash equilibrium respectively.

Quantal response equilibrium similarity classes. Figure A16 QRE(λ) rule similarity classes for $\lambda = 0.5036$.

Pareto efficiency similarity classes. Figure A17 shows the similarity classes for the Pareto efficiency rule.

Equal split and Soc-Max similarity classes. Figure A18 shows the similarity classes for the equal split rule; Figure A19 shows the similarity classes for the Soc-Max rule.

Equal split similarity classes. Finally, Figure A18 shows the similarity classes for the equal split rule;

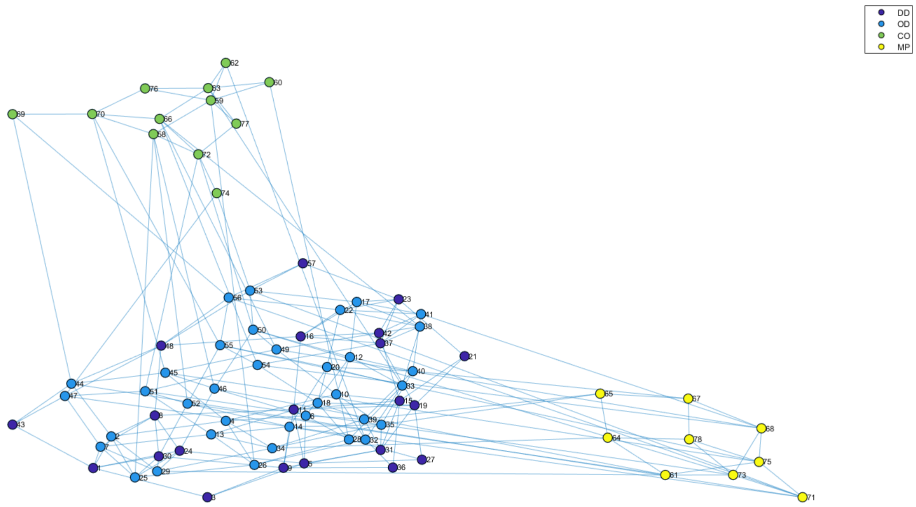


Figure A3: Graph of the Robinson-Goforth topology in G^* colored by types of games, with nodes representing the similarity classes according to Nash equilibrium and edges showing how they connect. Here, we also include the edges between games of the three different game classes. For ease of reading, they are not depicted in similar graphs.

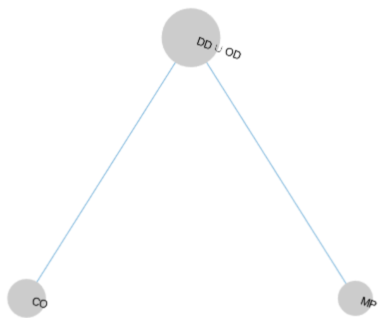


Figure A4: Graph of the Robinson-Goforth topology in G^*, with nodes representing the similarity classes according to Nash equilibrium and edges how they connect.

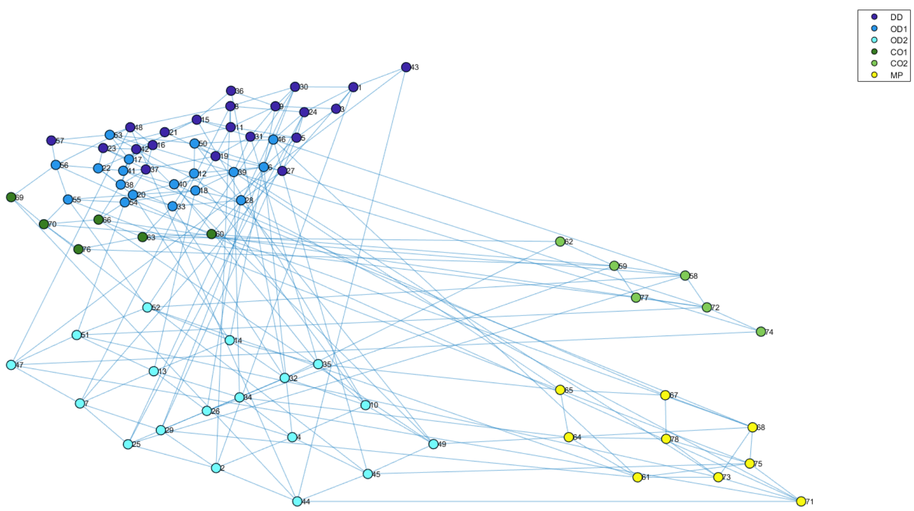


Figure A5: Graph of the Robinson-Goforth topology in G^* colored by similarity classes implied by Level-$k(\alpha)$ rules, $k = 1, \ldots, 5$.

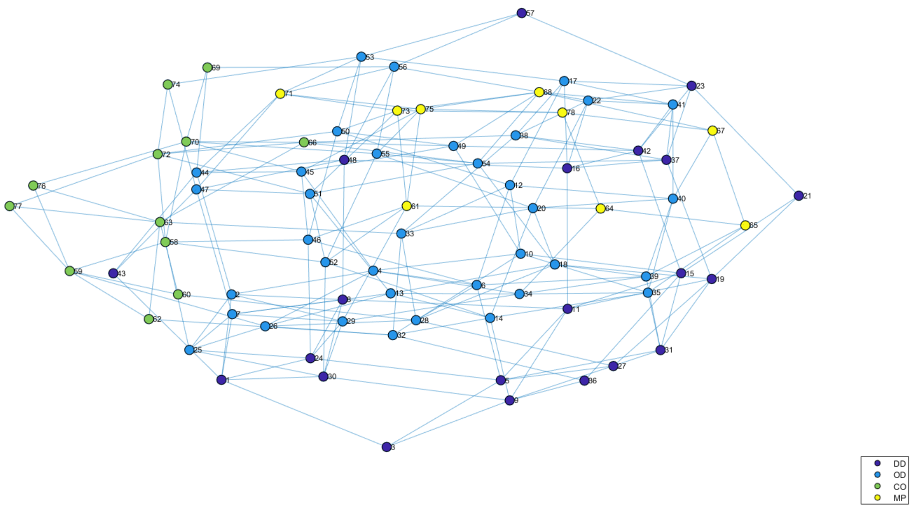


Figure A6: Graph of the similarity classes in G^* implied by the Level-1(α) rule. It does not separate any games. Each node is a game in G^* and a link is drawn between a pair of games if: (i) for some representation of the pair of games in G, both games are neighbors and (ii) both games belong to the same similarity class.

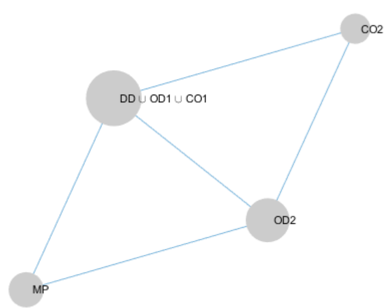


Figure A7: Graph of the Robinson-Goforth topology in G^*, with nodes representing the similarity classes according to Level-$k(\alpha)$ rules, $k = 1, \ldots, 5$, and edges how they connect.

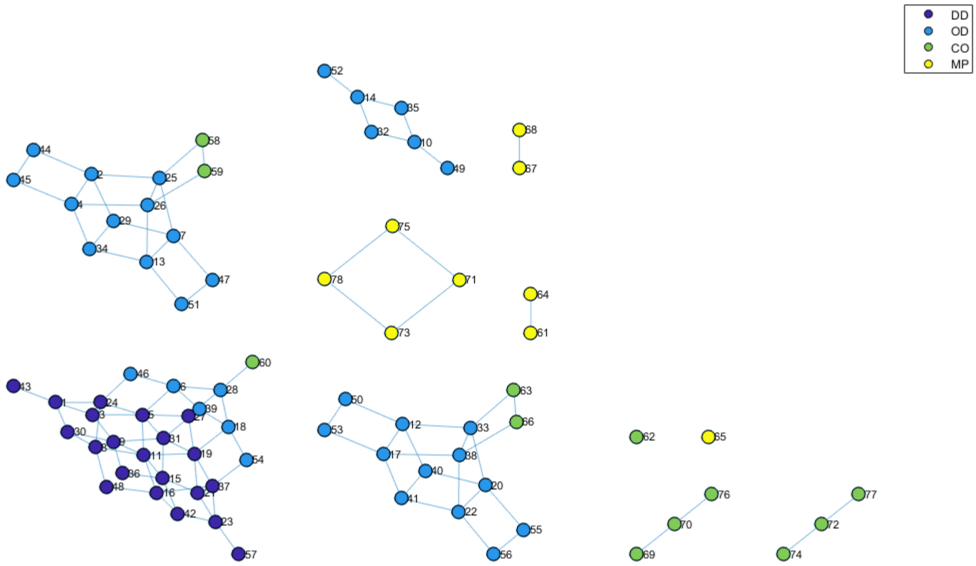


Figure A8: Graph of the similarity classes in G^* implied by the Level-k rule. Each node is a game in G^* and a link is drawn between a pair of games if: (i) for some representation of the pair of games in G, both games are neighbors and (ii) both games belong to the same similarity class.

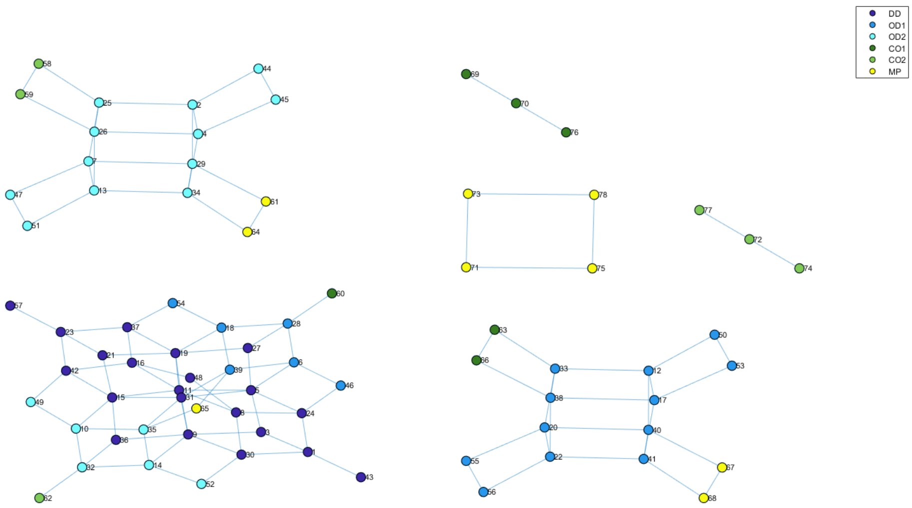


Figure A9: Graph of the similarity classes in G^* implied by the Level-1 rule. Each node is a game in G^* and a link is drawn between a pair of games if: (i) for some representation of the pair of games in G, both games are neighbors and (ii) both games belong to the same similarity class.

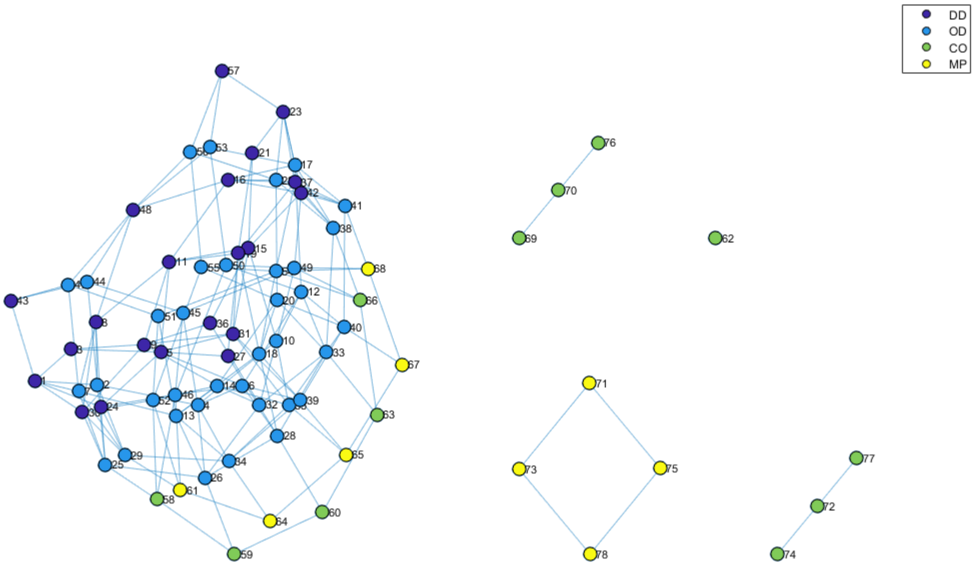


Figure A10: Graph of the similarity classes in G^* implied by the Level-2 rule. Each node is a game in G^* and a link is drawn between a pair of games if: (i) for some representation of the pair of games in G, both games are neighbors and (ii) both games belong to the same similarity class.

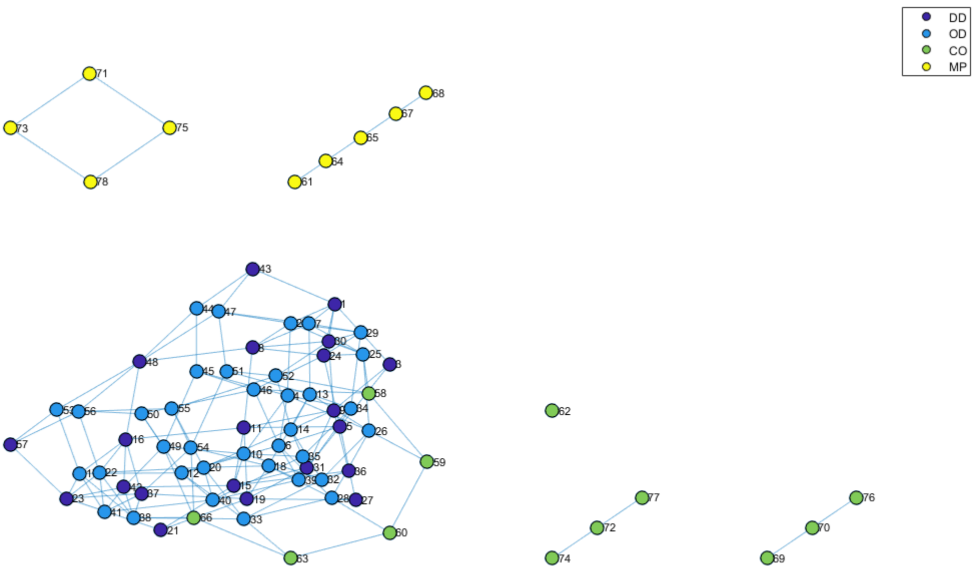


Figure A11: Graph of the similarity classes in G^* implied by the Level-4 rule. Each node is a game in G^* and a link is drawn between a pair of games if: (i) for some representation of the pair of games in G, both games are neighbors and (ii) both games belong to the same similarity class.

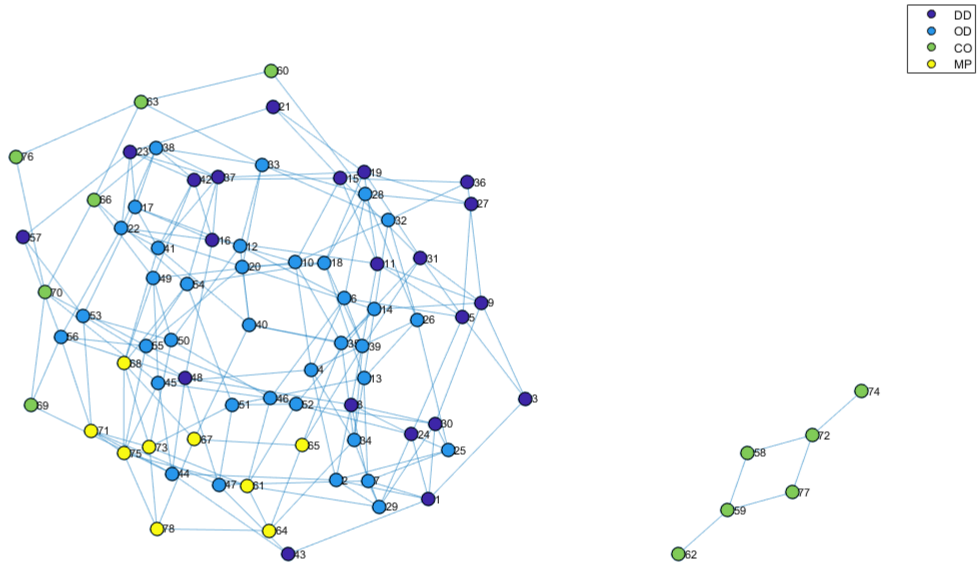


Figure A12: Graph of the similarity classes in G^* implied by the Level-5(α) rule. Each node is a game in G^* and a link is drawn between a pair of games if: (i) for some representation of the pair of games in G, both games are neighbors and (ii) both games belong to the same similarity class.

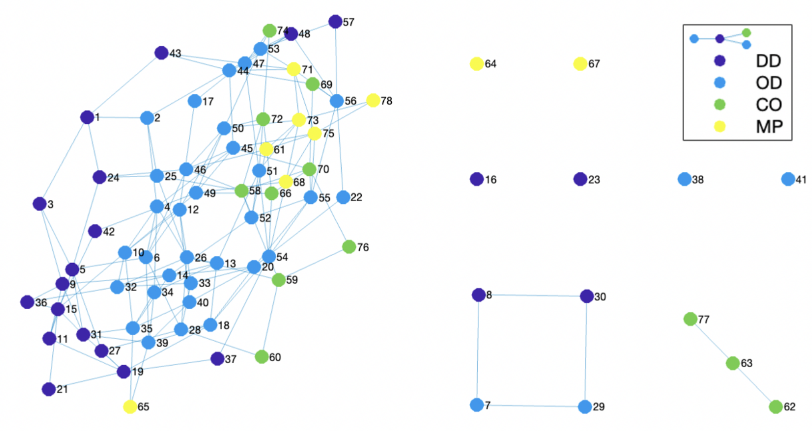


Figure A13: Graph of the similarity classes in G^* implied by self-favoring and other-favoring Near-Equal Split. Each node is a game in G^* and a link is drawn between a pair of games if: (i) for some representation of the pair of games in G, both games are neighbors and (ii) both games belong to the same similarity class.

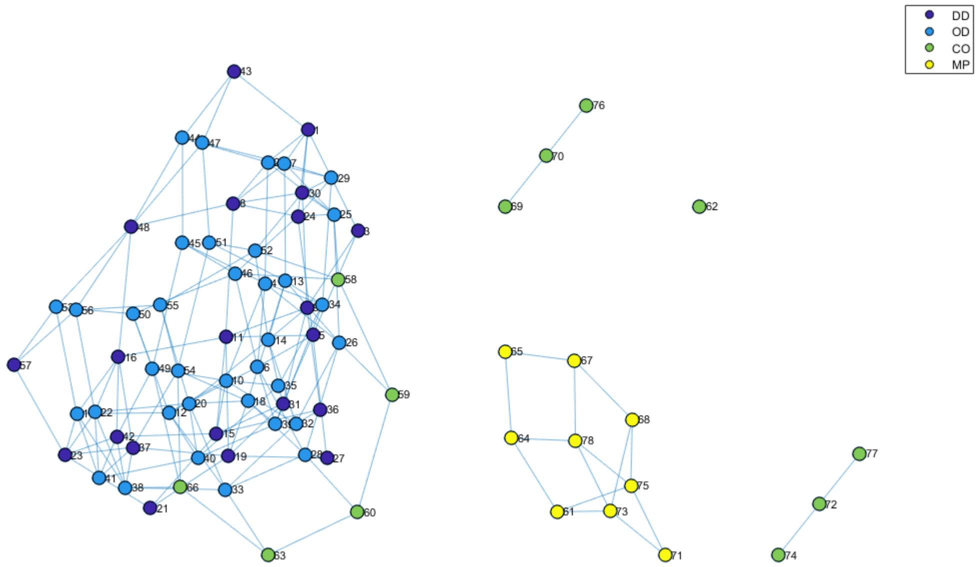


Figure A14: Graph of the similarity classes in G^* implied by Risk-Dominant Nash equilibrium. Each node is a game in G^* and a link is drawn between a pair of games if: (i) for some representation of the pair of games in G, both games are neighbors and (ii) both games belong to the same similarity class.

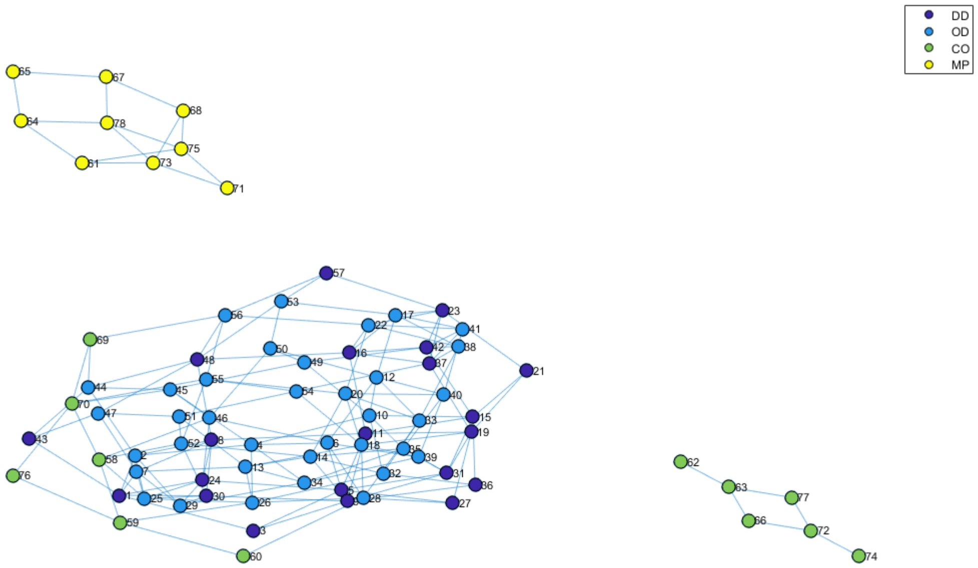


Figure A15: Graph of the similarity classes in G^* implied by Pareto-Dominant Nash equilibrium. Each node is a game in G^* and a link is drawn between a pair of games if: (i) for some representation of the pair of games in G, both games are neighbors and (ii) both games belong to the same similarity class.

Figure A16: Graph of the similarity classes in G^* implied by the QRE(λ) rule with $\lambda = 0.5036$. The value of λ is selected by 10-fold cross-validation to prevent overfitting. Each node is a game in G^* and a link is drawn between a pair of games if: (i) for some representation of the pair of games in G, both games are neighbors and (ii) both games belong to the same similarity class. In this case, two games are considered to have the same rule output if the difference in the QRE prediction is less than 5%.

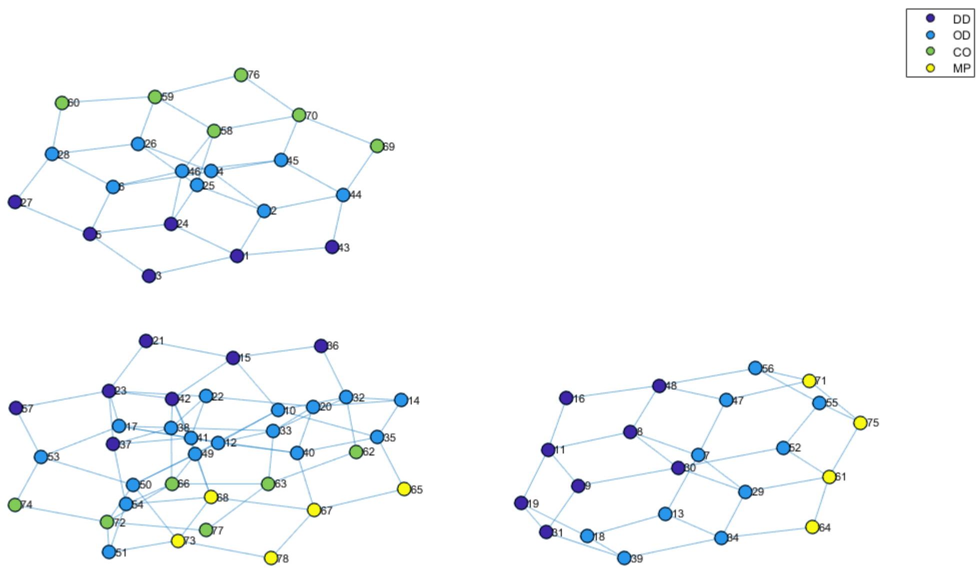


Figure A17: Graph of the similarity classes in G^* implied by Pareto Efficiency. Each node is a game in G^* and a link is drawn between a pair of games if: (i) for some representation of the pair of games in G, both games are neighbors and (ii) both games belong to the same similarity class.

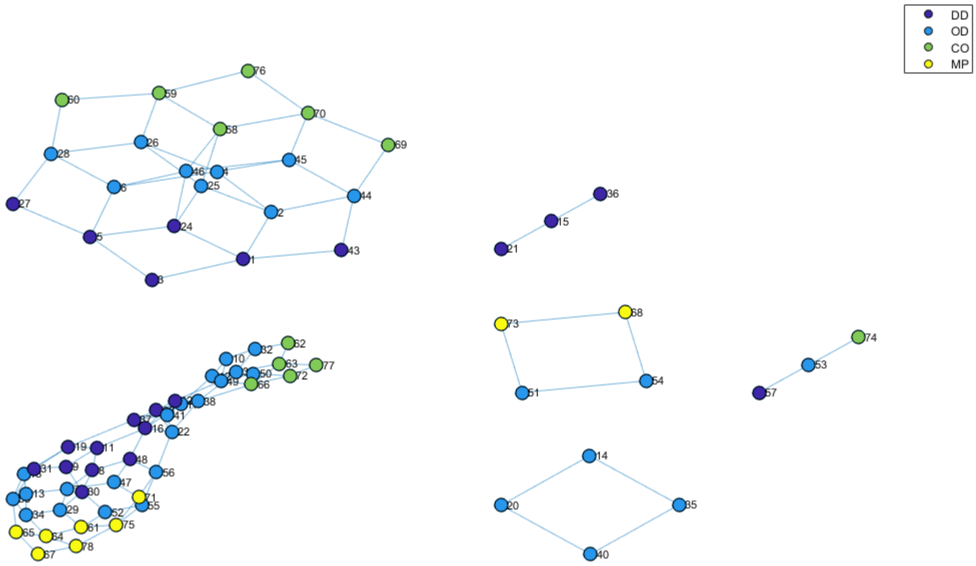


Figure A18: Graph of the similarity classes in G^* implied Equal Split. Each node is a game in G^* and a link is drawn between a pair of games if: (i) for some representation of the pair of games in G, both games are neighbors and (ii) both games belong to the same similarity class.

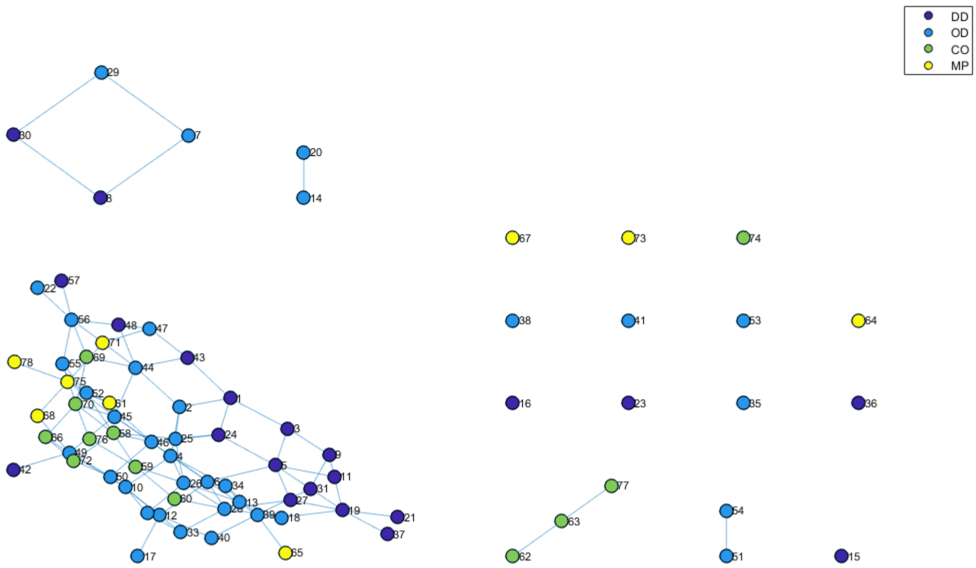


Figure A19: Graph of the similarity classes in G^* implied by the Soc-Max rule. Each node is a game in G^* and a link is drawn between a pair of games if: (i) for some representation of the pair of games in G, both games are neighbors and (ii) both games belong to the same similarity class.

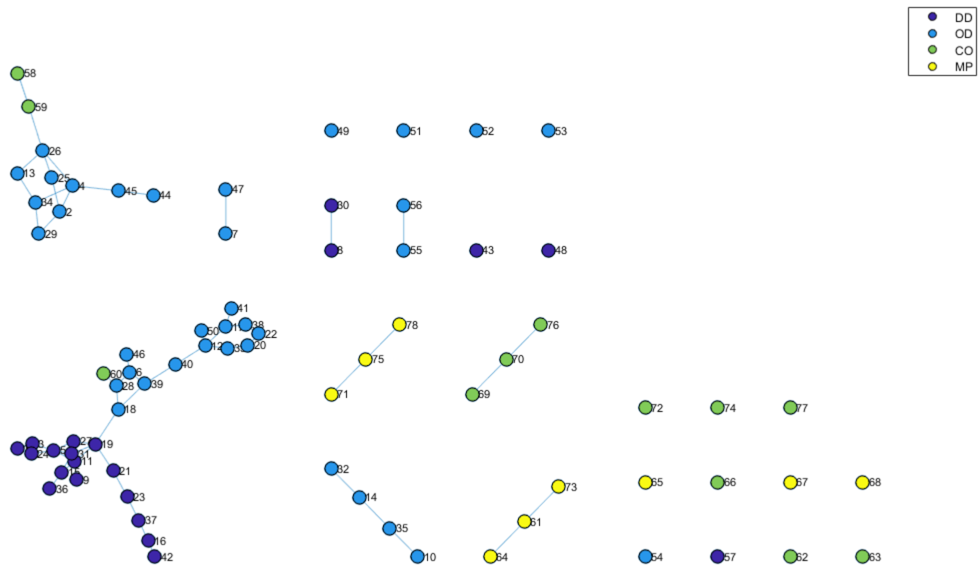


Figure A20: Graph of the empirical similarity classes in G^* from aggregate behavior of all subjects, for $\delta = 0\%$ at a 99.9% confidence level. Each node is a game in G^* and a link is drawn between a pair of games if: (i) for some representation of the pair of games in G, both games are neighbors and (ii) both games belong to the same similarity class.

B.3 Additional empirical similarity classes: All subjects

Empirical similarity classes for all subjects for $\delta = 0, 1$, and 2%. Figures A20, A21, and A22, shows the similarity classes for all subjects for confidence levels of, respectively, $\delta = 0\%$ at a 99.9%, $\delta = 1\%$ at a 99.9% and $\delta = 2\%$ at a 99.9%.

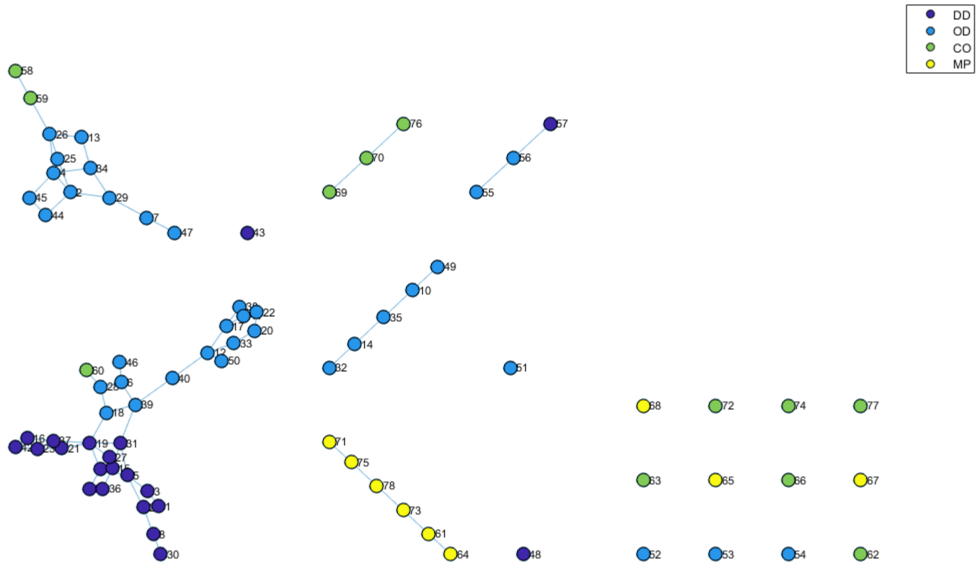


Figure A21: Graph of the empirical similarity classes in G^* from aggregate behavior of all subjects, for $\delta = 1\%$ at a 99.9% confidence level. Each node is a game in G^* and a link is drawn between a pair of games if: (i) for some representation of the pair of games in G, both games are neighbors and (ii) both games belong to the same similarity class.

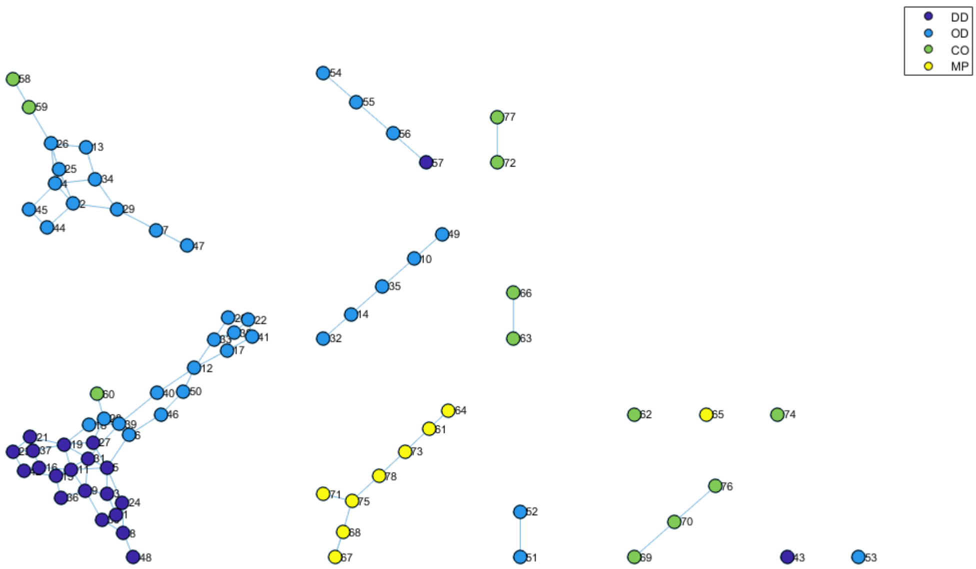


Figure A22: Graph of the empirical similarity classes in G^* from aggregate behavior of all subjects, for $\delta = 2\%$ at a 99.9% confidence level. Each node is a game in G^* and a link is drawn between a pair of games if: (i) for some representation of the pair of games in G, both games are neighbors and (ii) both games belong to the same similarity class.

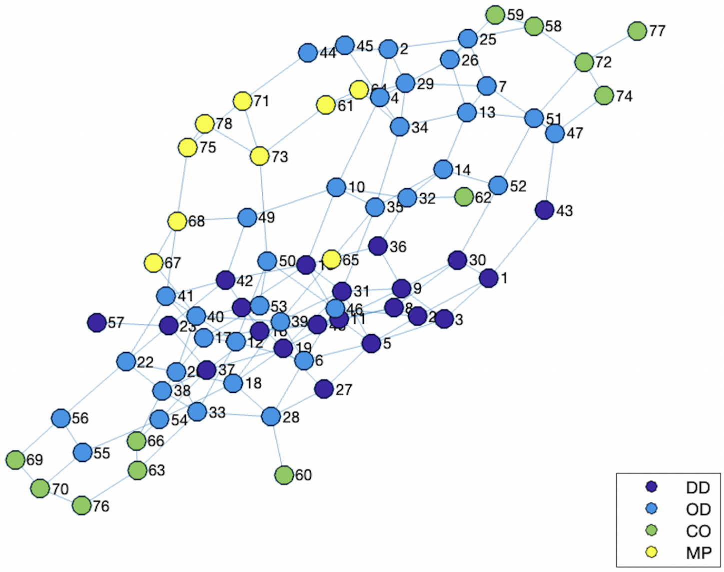


Figure A23: Graph of the empirical similarity classes in G^* from aggregate behavior of just those subjects that were identified as being level-1(α) types, for $(\Delta, p) = (5\%, 99.9\%)$.

B.4 Additional empirical similarity classes: Subgroups

Empirical similarity classes for L-1(α) subjects. Figure A23 shows the similarity classes for the subgroup of subjects classified as L1-(α) types.

Empirical similarity classes for L-k(α) subjects, $k = 2, \ldots, 5$. Figure A24 shows the similarity classes for the subgroup of subjects classified as L1-(α) types.

Empirical similarity classes for near-equal split subjects. Figure A25 shows the similarity classes for the subgroup of subjects classified as rNES or cNES types.

B.5 Additional figures on individual behavior

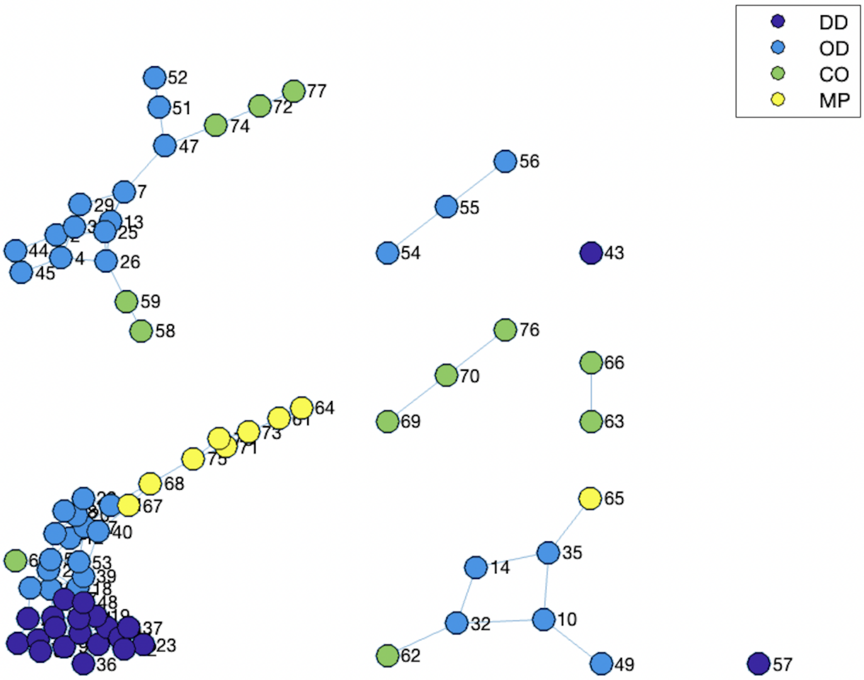


Figure A24: Graph of the empirical similarity classes in G^* from aggregate behavior of just those subjects that were identified as being level-$k(\alpha)$ types, for $k = 2, \ldots, 5$, and for $(\Delta, p) = (5\%, 99.9\%)$.

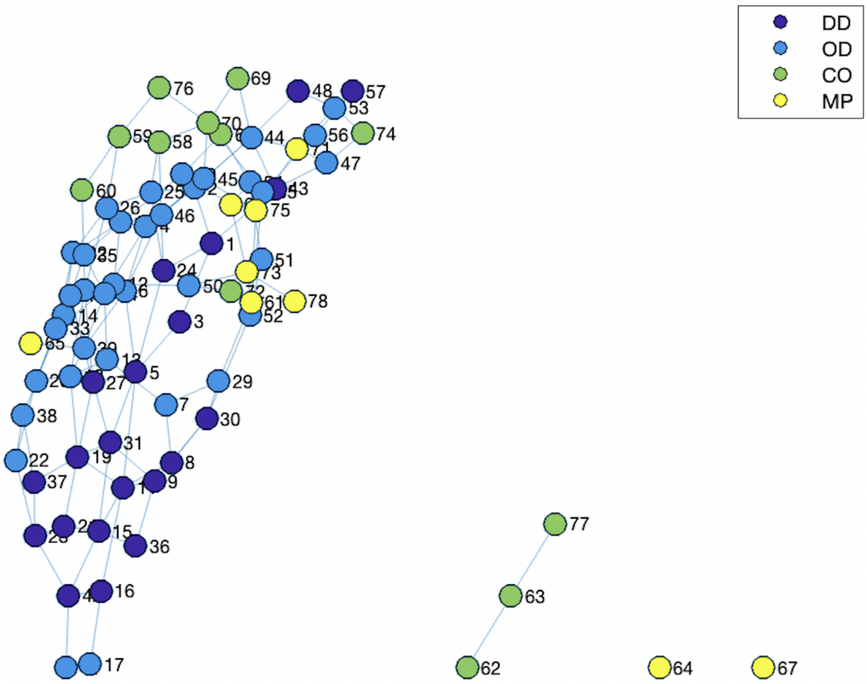


Figure A25: Graph of the empirical similarity classes in G^* from aggregate behavior of just those subjects that were identified as being cNES or rNES types, for $(\Delta, p) = (5\%, 99.9\%)$.

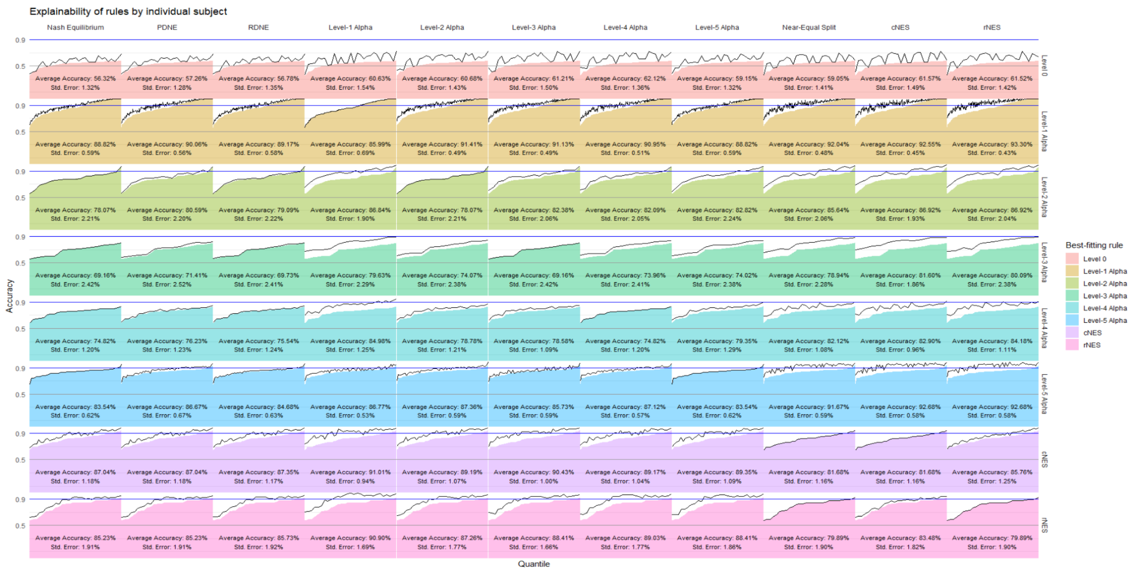


Figure A26: The quantile function of the accuracy of each subject's best-fitting rule (area in color) plus the added accuracy of other rules in games where the best-fitting rule predicts incorrectly. The rows correspond to the groups of subjects according to their best-fitting rule. The columns correspond to the rule used to predict games where the best-fitting rule predicts incorrectly. This figure reports sNES and oNES as rNES and cNES respectively.

Primary/Sec.	cNES	Level-1	Level-1 Alpha	Level-2	Level-2 Alpha	Level-3 Alpha	Level-4 Alpha	Nash Equilibrium	Near-Equal Split	PDNE	rNES	Soc-Max
cNES	0	0	25	0	4	11	0	1	0	0	1	0
Level-1 Alpha	41	0	0	2	20	13	18	5	21	1	75	17
Level-2	1	0	0	0	0	0	0	0	0	0	0	0
Level-2 Alpha	7	1	7	0	0	0	0	0	0	0	4	0
Level-3 Alpha	7	0	2	0	0	0	0	0	0	0	2	1
Level-4 Alpha	4	0	16	0	0	0	0	0	0	0	7	0
Level-5 Alpha	34	0	3	0	0	1	0	0	1	0	24	2
Level 0	14	0	4	0	0	4	5	0	0	0	2	0
rNES	0	0	24	0	0	3	2	0	0	0	0	0

Figure A27: Numbers of subjects by best-fitting secondary rule (columns), for each best-fitting primary rule (rows).

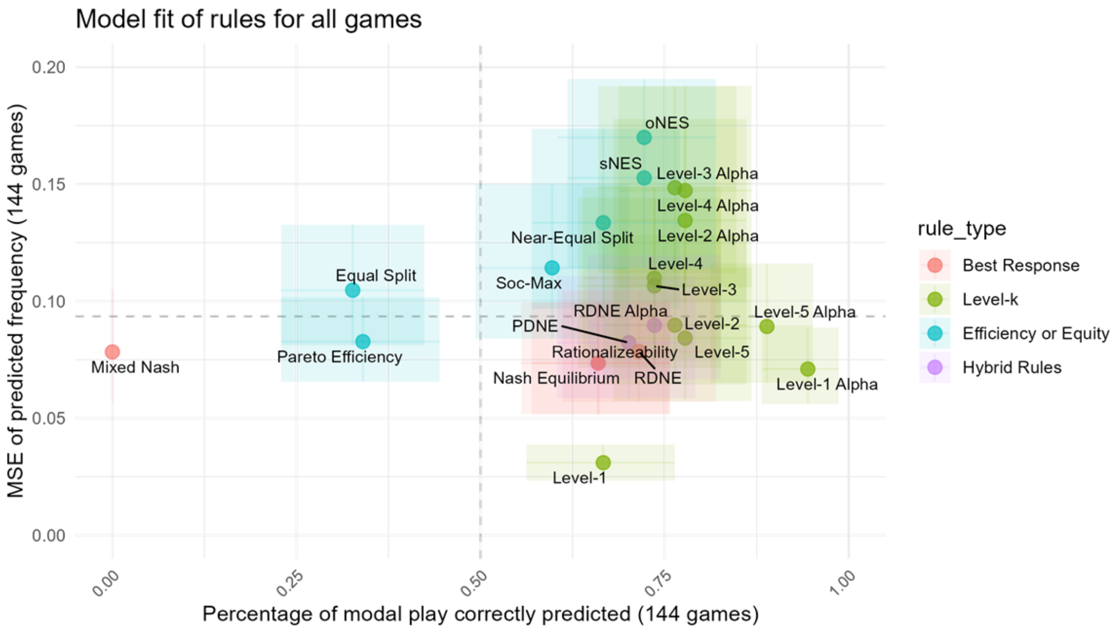


Figure A28: Model fit for behavior rules across all 144 perspectives. The x-axis corresponds to accuracy in modal play whereas the y-axis corresponds to MSE. The rectangles correspond to 99% confidence intervals of each statistic.

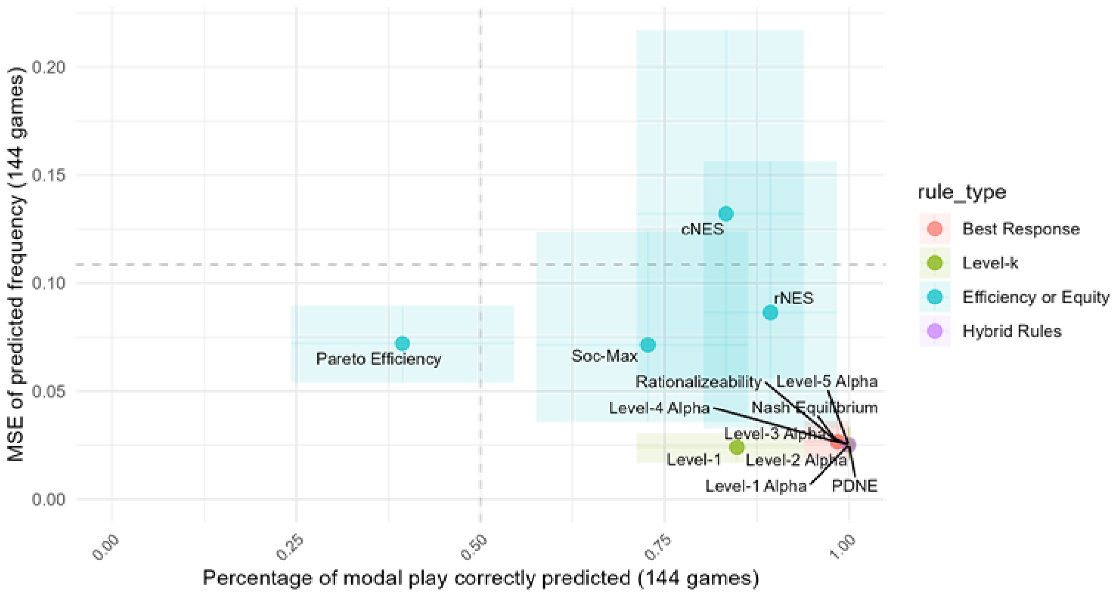


Figure A29: Model fit for behavior rules across the perspectives in the One-outcome empirical similarity class. The x-axis corresponds to accuracy in modal play whereas the y-axis corresponds to MSE. The rectangles correspond to 99% confidence intervals of each statistic.

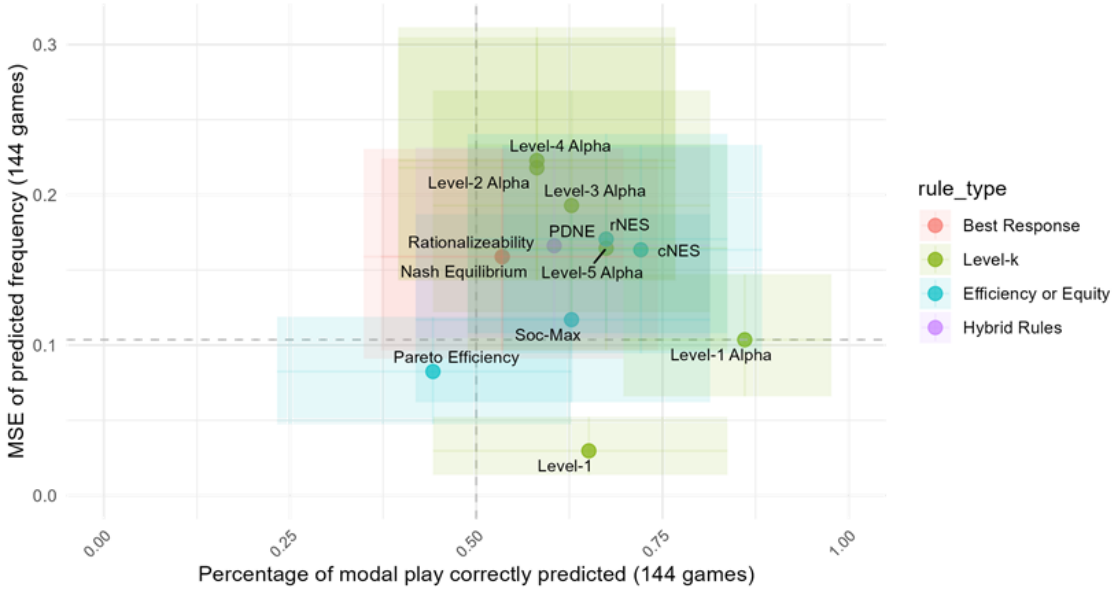


Figure A30: Model fit for behavior rules across the perspectives in the two outcome empirical similarity class. The x-axis corresponds to accuracy in modal play whereas the y-axis corresponds to MSE. The rectangles correspond to 99% confidence intervals of each statistic.

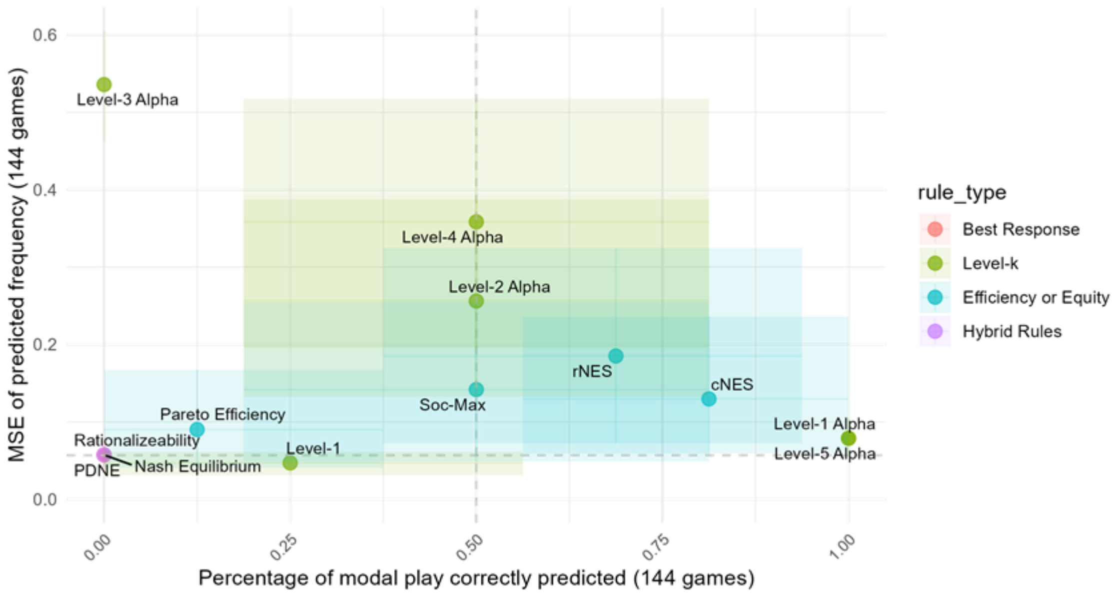


Figure A31: Model fit for behavior rules across the perspectives in the matching pennies empirical similarity class. The x-axis corresponds to accuracy in modal play whereas the y-axis corresponds to MSE. The rectangles correspond to 99% confidence intervals of each statistic.

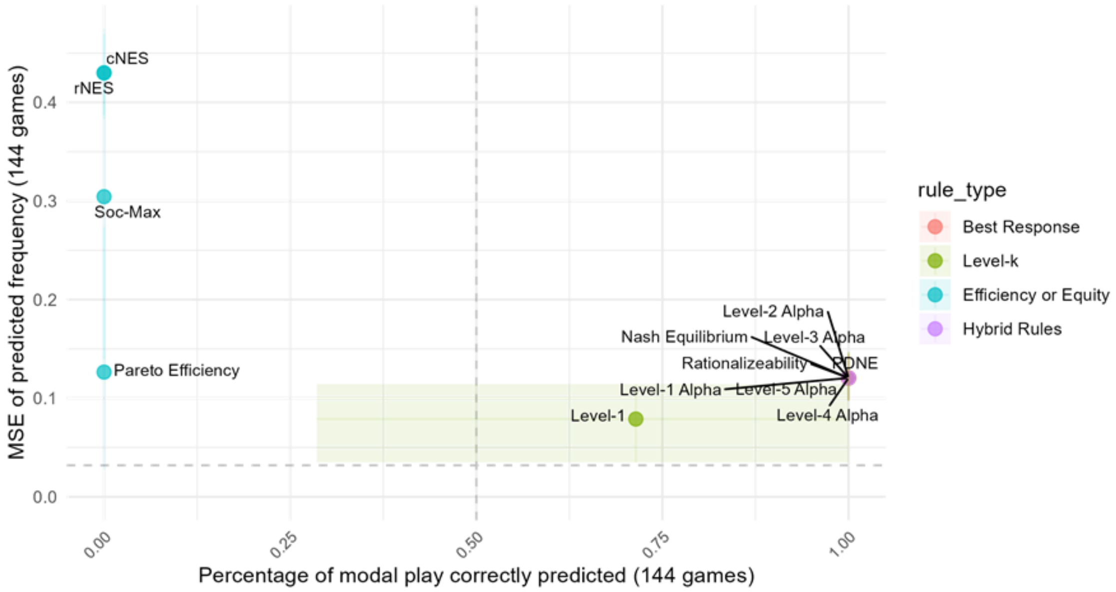


Figure A32: Model fit for behavior rules across the perspectives in the Prisoner's Dilemma empirical similarity class. The x-axis corresponds to accuracy in modal play whereas the y-axis corresponds to MSE. The rectangles correspond to 99% confidence intervals of each statistic.

C Comparison of the 44-Games Experiment and the All-Games Experiment

Prior to the experimental design described in the main paper, we implemented a preliminary design featuring sequences of 44 games. These sequences were generated quasi-randomly to address several objectives simultaneously. First, to ensure sufficient power for detecting differences in games predicted to have noisy outcomes (where subjects were equally likely to choose A or B), we oversampled CO and MP games relative to DD and OD games. Second, to span all edges in the Robinson-Goforth topology and ensure enough power to compare behavior across neighboring games, we assigned pairs of neighboring games randomly to each sequence until all pairs had sufficient coverage. Finally, we filled the remaining slots in each sequence with underrepresented game classes to guarantee representation for every game. For example, if a sequence contained primarily DD games after assigning neighboring pairs, CO and MP games were added to balance the game types encountered by each subject.

After constructing the sequences, we treated them similarly to the sequences in the final design: the order of games was randomized, and the labels for games were also randomized. Notably, the sequences in this preliminary design were much shorter, with each subject playing only 44 games. A total of 30 sequences were created, each consisting of 44 games, and we recruited 960 subjects from Prolific, assigning 32 subjects per sequence. Subjects received a show-up fee of 1 GBP and were compensated at a rate of 0.40 GBP per average number of points earned across three randomly selected games.[32]

[32]This compensation structure technically provided higher stakes per hour compared to the final design.

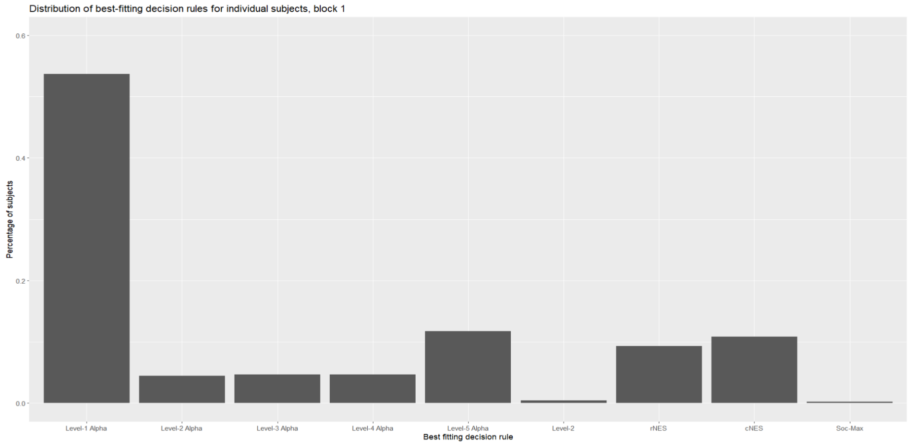


Figure A33: The distribution of best-fitting rules for our 450 subjects, using only the first 48 perspectives that subjects played.

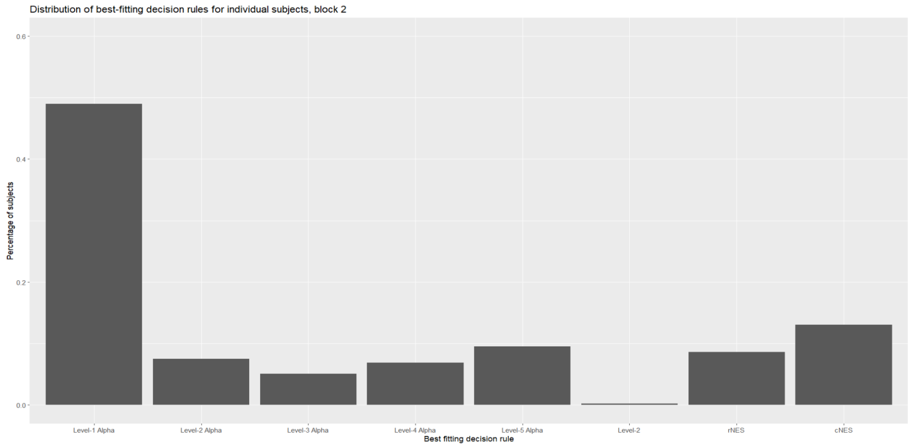


Figure A34: The distribution of best-fitting rules for our 450 subjects, using only the 49th to 95th perspectives that subjects played.

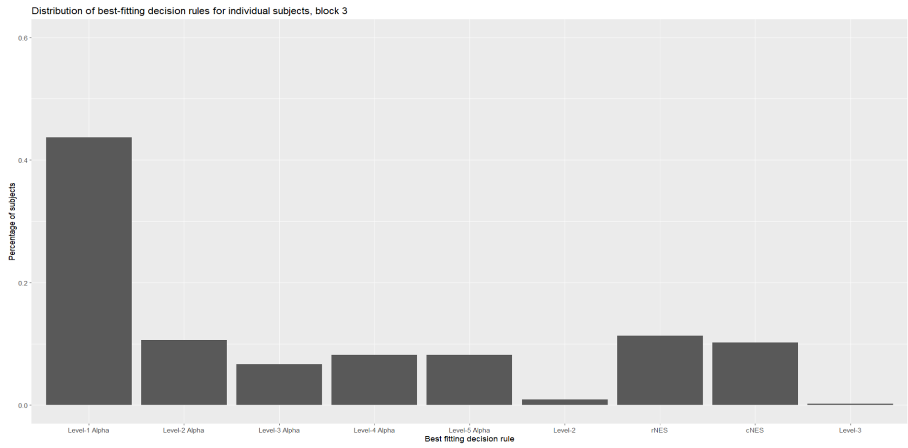


Figure A35: The distribution of best-fitting rules for our 450 subjects, using only the last 48 perspectives that subjects played.

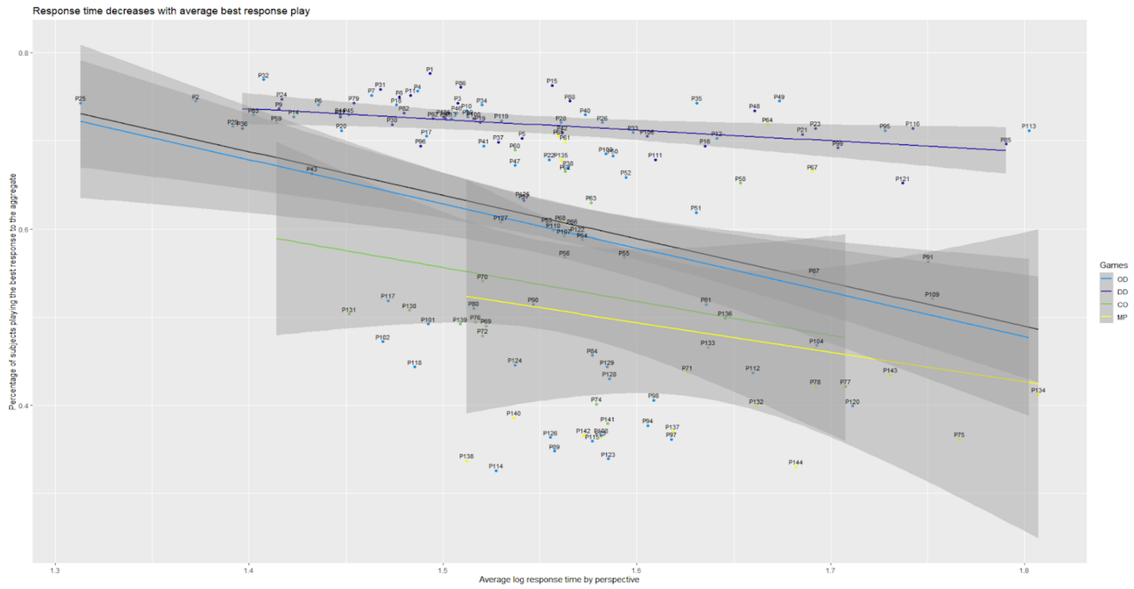
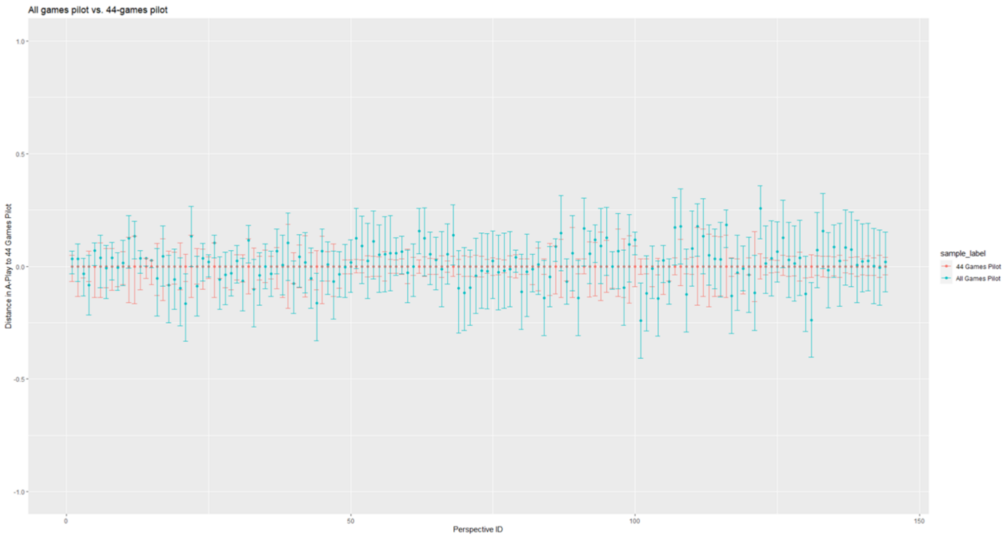


Figure A36: The relationship between log response time by perspective and the percentage of subjects playing the best response to the frequency distribution for each game.

The average completion time for this experiment was significantly shorter, at around 8 minutes. The qualitative results obtained from this preliminary design align closely with those from the final all-games experiment. To test this, we calculated the 90% confidence intervals for the proportion of subjects choosing A for each perspective in both experiments. The results are summarized in Figure A37. Across the 144 perspectives, the intervals from the 44-games experiment include the point estimates of the all-games experiment in 70% (101 out of 144) of perspectives. Although this is lower than the theoretical expectation of 90%, we observed no significant skew in the direction of the bias.

D Table of games in G^*

In the 4-page table below, the column denoted by "type" distinguishes four broad types of games: games with one-sided strict dominance (OD), games with two-sided strict dominance (DD), matching pennies type games (MP), and coordination type games (CO). However, it is useful to further refine OD and CO depending on whether the Level-$k(\alpha)$ rules select just one or more than One-outcome. Hence we have two types of games in OD: ones where the Level-$k(\alpha)$ rules select a single outcome (OD1), and ones where they select Two-outcomes (OD2). Similarly, there are two types of coordination games: ones where Level-$k(\alpha)$ rules select a single NE (CO1), and ones ones where Level-$k(\alpha)$ rules span all outcomes and hence both NE (CO2).



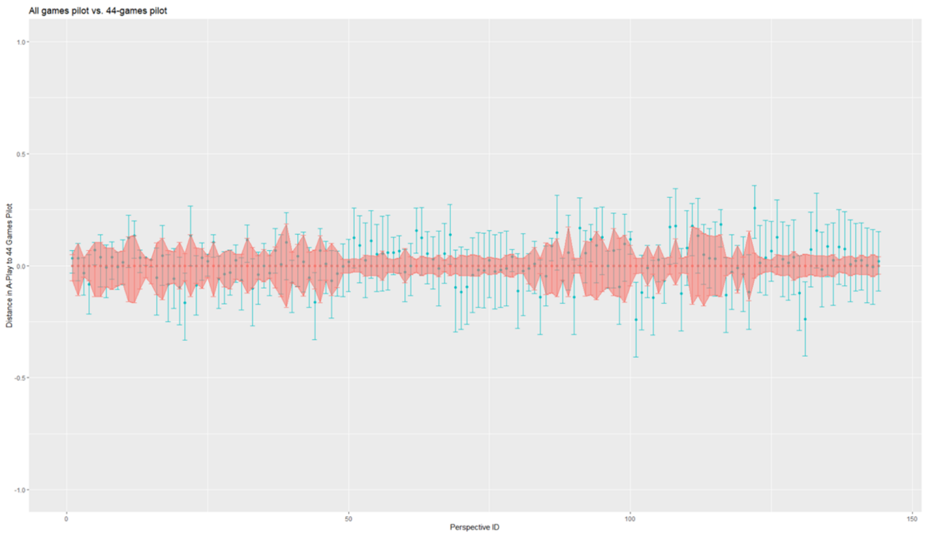


Figure A37: Difference in A-Play between the 44 games pilot and the data from the final design. Error bars correspond to 90% confidence interval.

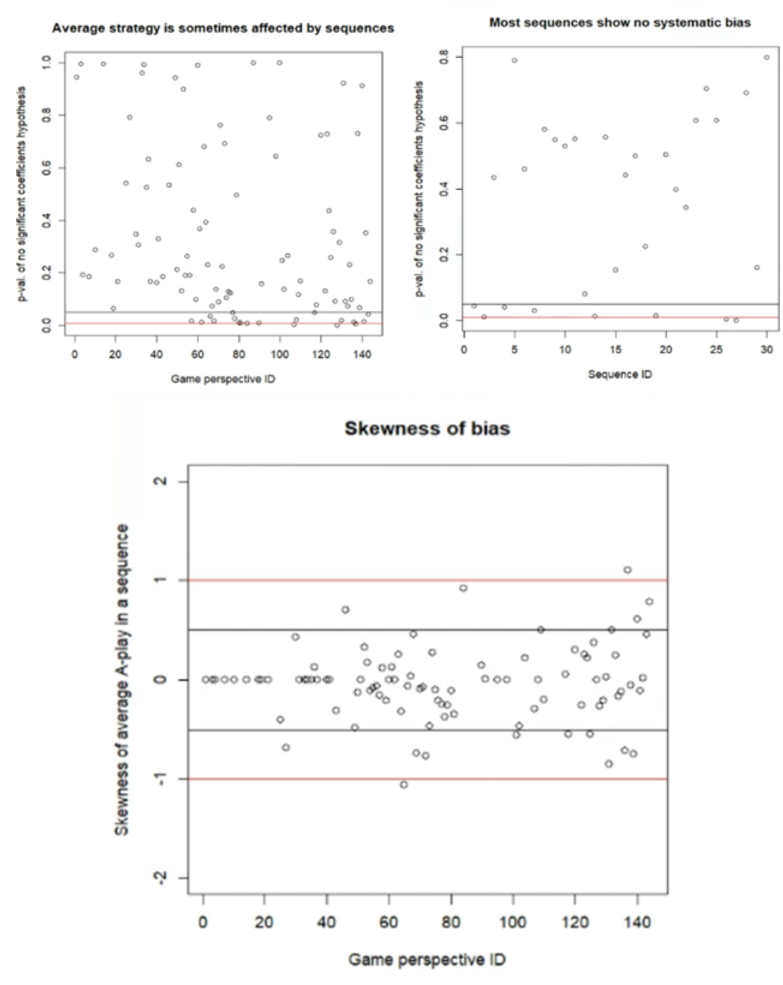


Figure A38: Summary of sequence effects in the all games experiment.

E. Table of games in G^*

Game id	Payoff matrix				Bruns id	L$k(\alpha)$ type	Emp. similarity class
1	4	4	3	3	11,1	DD	A1
	2	2	1	1	12,1		
2	4	4	3	3	7,2	OD2	A2
	2	1	1	2	11,6		
3	4	4	3	2	11,2	DD	A1
	2	3	1	1			
4	4	4	3	2	8,2	OD2	A2
	2	1	1	3	11,5		
5	4	4	3	1	10,2	DD	A1
	2	3	1	2	11,3		
6	4	4	3	1	9,2	OD1	A1
	2	2	1	3	11,4		
7	4	3	3	4	1,2	OD2	A2
	2	2	1	1	11,12		
8	4	3	3	4	6,2	DD	A1
	2	1	1	2	11,7		
9	4	3	3	2	5,3	DD	A1
	2	4	1	1	10,8		
10	4	3	3	2	3,8	OD2	A2
	2	1	1	4	5,1		
11	4	3	3	1	5,2	DD	A1
	2	4	1	2	11,8		
12	4	3	3	1	2,8	OD1	A1
	2	2	1	4	5,1		
13	4	2	3	4	2,2	OD2	A2
	2	3	1	1	11,11		
14	4	2	3	3	5,4	OD2	A2
	2	4	1	1	9,8		
15	4	2	3	3	4,8	DD	A1
	2	1	1	4	5,9		
16	4	2	3	1	5,1	DD	A1
	2	4	1	3	12,8		
17	4	2	3	1	1,8	OD1	A1
	2	3	1	4	5,12		
18	4	1	3	4	3,2	OD1	A1
	2	3	1	2	11,10		
19	4	1	3	4	4,2	DD	A1
	2	2	1	3	11,9		
20	4	1	3	3	5,5	OD1	A1
	2	4	1	2	8,8		

E. Table of Games in G^* (cont.)

Game id	Payoff matrix	Bruns id	L$k(\alpha)$ type	Emp. similarity class
21	4 1 3 3 2 2 1 4	5,8	DD	A1
22	4 1 3 2 2 4 1 3	5,6 7,8	OD1	A1
23	4 1 3 2 2 3 1 4	5,7 6,8	DD	A1
24	4 4 3 3 1 2 2 1	10,1 12,3	DD	A1
25	4 4 3 3 1 1 2 2	7,3 10,6	OD2	A2
26	4 4 3 2 1 1 2 3	8,3 10,5	OD2	A2
27	4 4 3 1 1 3 2 2	10,3	DD	A1
28	4 4 3 1 1 2 2 3	9,3 10,4	OD1	A1
29	4 3 3 4 1 2 2 1	1,3 10,12	OD2	A2
30	4 3 3 4 1 1 2 2	6,3 10,7	DD	A1
31	4 3 3 2 1 4 2 1	4,3 10,9	DD	A1
32	4 3 3 2 1 1 2 4	3,9 4,10	OD2	A2
33	4 3 3 1 1 2 2 4	2,9 4,11	OD1	A1
34	4 2 3 4 1 3 2 1	2,3 10,11	OD2	A2
35	4 2 3 3 1 4 2 1	4,4 9,9	OD2	A2
36	4 2 3 3 1 1 2 4	4,9	DD	A1
37	4 2 3 1 1 4 2 3	4,1 12,9	DD	A1
38	4 2 3 1 1 3 2 4	1,9 4,12	OD1	A1
39	4 1 3 4 1 3 2 2	3,3 10,10	OD1	A1
40	4 1 3 3 1 4 2 2	4,5 8,9	OD1	A1

E. Table of Games in G^* (cont.)

Game id	Payoff matrix				Bruns id	L$k(\alpha)$ type	Emp. similarity class
41	4	1	3	2	4,6	OD1	A1
	1	4	2	3	7,9		
42	4	1	3	2	4,7	DD	A1
	1	3	2	4	6,9		
43	4	4	2	3	12,1	DD	A1
	3	2	1	1			
44	4	4	2	3	7,1	OD2	A2
	3	1	1	2	12,6		
45	4	4	2	2	8,1	OD2	A2
	3	1	1	3	12,5		
46	4	4	2	1	9,1	OD1	A1
	3	2	1	3	12,4		
47	4	3	2	4	1,1	OD2	A2
	3	2	1	1	12,12		
48	4	3	2	4	6,1	DD	A1
	3	1	1	2	12,7		
49	4	3	2	2	3,7	OD2	A2
	3	1	1	4	6,10		
50	4	3	2	1	2,7	OD1	A1
	3	2	1	4	6,11		
51	4	2	2	4	2,1	OD2	A2
	3	3	1	1	12,11		
52	4	2	2	3	6,4	OD2	A2
	3	4	1	1	9,7		
53	4	2	2	1	1,7	OD1	A1
	3	3	1	4	6,12		
54	4	1	2	4	3,1	OD1	B1
	3	3	1	2	12,10		
55	4	1	2	3	6,5	OD1	B1
	3	4	1	2	8,7		
56	4	1	2	2	6,6	OD1	B1
	3	4	1	3	7,7		
57	4	1	2	2	6,7	DD	B1
	3	3	1	4			
58	4	4	2	3	7,4	CO2	A2
	1	1	3	2	9,6		
59	4	4	2	2	8,4	CO2	A2
	1	1	3	3	9,5		
60	4	4	2	1	9,4	CO1	A1
	1	2	3	3			

E. Table of Games in G^* (cont.)

Game id	Payoff matrix				Bruns id	L$k(\alpha)$ type	Emp. similarity class
61	4	3	2	4	1,4	MP	A4
	1	2	3	1	9,12		
62	4	3	2	2	3,10	CO2	B4
	1	1	3	4			
63	4	3	2	1	2,10	CO1	B3
	1	2	3	4	3,11		
64	4	2	2	4	2,4	MP	A4
	1	3	3	1	9,11		
65	4	2	2	3	3,4	MP	A2
	1	4	3	1	9,10		
66	4	2	2	1	1,10	CO1	B3
	1	3	3	4	3,12		
67	4	1	2	3	3,5	MP	A4
	1	4	3	2	8,10		
68	4	1	2	2	3,6	MP	A4
	1	4	3	3	7,10		
69	4	4	1	3	7,6	CO1	B2
	3	1	2	2			
70	4	4	1	2	7,5	CO1	B2
	3	1	2	3	8,6		
71	4	3	1	4	1,6	MP	A4
	3	2	2	1	7,12		
72	4	3	1	1	1,11	CO2	A3
	3	2	2	4	2,12		
73	4	2	1	4	2,6	MP	A4
	3	3	2	1	7,11		
74	4	2	1	1	1,12	CO2	A2
	3	3	2	4			
75	4	1	1	3	1,5	MP	A4
	3	4	2	2	8,12		
76	4	4	1	2	8,5	CO1	B2
	2	1	3	3			
77	4	3	1	1	2,11	CO2	A3
	2	2	3	4			
78	4	2	1	4	2,5	MP	A4
	2	3	3	1	8,11		