Universitat Pompeu Fabra Barcelona
Department of Economics and Business

Economics Working Paper Series

Working Paper No. 1942

Understanding human behavior via similarity: A geometric and behavioral Rules-based approach to games

Amil Camilo Moore, Fabrizio Germano, and Rosemarie Nagel

March 2026

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

Amil Camilo Moore† Fabrizio Germano‡ Rosemarie Nagel§

March 27, 2026

Abstract

We study similarity in the complete set of one-shot 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 (e.g., Nash equilibrium, level-k reasoning) and experimental data. This produces a partition of the games into (theoretical or empirical) *similarity classes*. We run a large-scale experiment in which each subject plays all 78 games within our class without feedback. We find that empirically inferred similarity classes diverge sharply from those predicted by Nash equilibrium and dominance reasoning. Instead, the empirical similarity classes align closely with the theoretical classes of 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 variant ($0 \leq k \leq 5$) or a fairness and efficiency-based heuristic. The main insights extend to strategic settings beyond our 2×2 games.

Keywords. Similarity of games; level-k reasoning; equity and efficiency; experiments; topology of games. **JEL Classification**. C52, C70, C72, C81, C90, C93, D91.

*We thank Larbi Alaoui, Jose Apesteguia, Elia Benveniste, Ayah Bohsali, Luke Boosey, Antonio Cabrales, Victoria Cirlot, Vince Crawford, John Duffy, Pia Ennuschat, Ido Erev, Vinicius Ferraz, Dan Friedman, Evan Friedman, Alexander Frug, Malachy Gavan, Ben Golub, Duarte Gonçalves, Glenn Harrison, R. Mark Isaac, Gaël Le Mens, Annie Liang, Raquel Lorenzo, Antonio Penta, Yaroslav Rosokha, Al Roth, Andrea Salvanti, Marta Serra, Patrick Sewell, Colin Sullivan, Elias Tsakas, Robert Wojciechowski, and audiences in seminars and conferences in Barcelona, Bielefeld, Budapest, Lancaster, London, New Orleans, Stony Brook, Tallahassee, 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, and BSE, 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). Game theorists tend to classify games according to the properties of the games' equilibria and other logical features of the games. Despite the centrality of similarity to cognition and individual decision making, the literature has yet to provide, to the best of our knowledge, a systematic study of similarity across a wide variety of games based on either bounded rationality or how people actually play games.[1]

Addressing this gap requires a notion of similarity that can translate the distinctions game theorists tend to make into a quantifiable structure of behavior. To formalize this, we adopt Germano's (2006) geometric framework, which groups games by their neighborhood structure and by the theoretical continuity of choices, as prescribed or actually taken across games. We extend this approach by applying it, for the first time, to behavioral rules and experimental data. Specifically, for any rule or list of choices defined on a given space of games (e.g. Nash equilibrium actions over 2×2 games), we regard two games as similar if there exists a path of neighboring games such that the choices made along the path of games remain continuous. When applying this approach to experimental data, our notion of continuity relies on the differences in the choice frequencies between neighboring games being sufficiently small relative to a statistical threshold. The discontinuities of the choices partition the space of games into a finite number of connected components, which we refer to as similarity classes. These similarity classes naturally provide a classification of games.

We study similarity between games in the most basic strategic environment: the full set of one-shot 2×2 games (meaning, two players choose between two actions) in which payoffs are drawn without replacement from the set $\{1, 2, 3, 4\}$.[2] We draw especially on the work by Rapoport, Guyer, and Gordon (1978), Robinson and Goforth (2005), Shubik (2012), and Bruns (2015).[3] Despite its simplicity, this environment serves as a microcosm of strategic interaction, capturing fundamental structures underlying many different economic situations. For example, it includes many of the most famous normal-form games, including the Prisoner's Dilemma, Battle of the

[1] For a discussion of similarity in behavioral game theory, see the related literature section further below.

[2] While Germano (2006) studies spaces of finite games as Euclidean spaces with corresponding notions of continuity of correspondences (see Aliprantis and Border, 2006), in our case, two games are neighbors if they are closest in the discrete subspace, 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.

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

Sexes, Stag Hunt, Chicken, and Matching Pennies. By studying this complete domain of games, we can analyze how small changes in payoffs predict discontinuous shifts in behavior, and relate these shifts to game-theoretic principles and behavioral rules. At the same time, the simplicity of this domain also makes the number of playable games small enough to implement an experiment in which we can consistently measure how subjects behave. As a byproduct of our analysis, we uncover understudied games such as asymmetric variants of the Prisoner's Dilemma, Battle of the Sexes, and Stag Hunt.

One of our contributions is that we offer foundational insights that we hope will be applicable to more general strategic settings. [4] For instance, we find that the aggregate behavior of our subjects organizes this domain of games into empirical similarity classes which are best matched by the similarity classes implied by a risk-averse variant of level-k reasoning, level-$k(\alpha)$.[5] However, the differences between the similarity classes from the aggregate behavior and the similarity classes from level-$k(\alpha)$ are driven by disagreements between level-$k(\alpha)$ and a rule that induces fair and efficient outcomes that we refer to as near-equal split (NES).[6]

More concretely, the level-$k(\alpha)$ rule partitions the games into four theoretical similarity classes based on the steps of reasoning required to reach a Nash equilibrium. These classes align closely with the first four empirical classes identified in our data (hereafter A1–A4). The first level-$k(\alpha)$ class (A1) consists of games in which the level-$k(\alpha)$ prediction converges already to a Nash equilibrium at $k = 1$. This includes the double-sided dominance games (DD) where both players have a dominant strategy, some one-sided dominance games (OD), where only one player has a dominant strategy, and some coordination games (CO). The second class (A2) consists of games in which convergence to the unique NE requires two steps for one player ($k = 2$) and includes the remaining OD games. The third and fourth classes (A3 and A4) contain CO games in which the L$k(\alpha)$ rules lead to off-equilibrium outcomes in coordination and games in which there is no pure-strategy equilibrium (or Matching Pennies type (MP) games). While these four classes fit the aggregate data well in most games, they diverge in games in which L1(α) or NE do not align with NES. Because of such empirical discontinuities, four additional classes (called B1–B4) arise in the data. These "breakaway" classes identify games where social preferences or efficiency concerns override simple level-$k(\alpha)$ reasoning for a sufficient number of subjects, most notably in Prisoner's Dilemma- (B1), Stag Hunt- (B2), and Battle of the Sexes-type (B3 and B4) environments. This empirical richness stands in sharp contrast to the coarser similarity classes provided by Nash Equilibrium, which groups games into only three classes (DD $\cup$ OD, CO, and MP) and Strict

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

[5] Strategies consistent with level-k reasoning are iterated best responses to uniform random play: concretely, level-1 is a best response to uniform random play (or level-0), level-2 is a best response to an opponent that plays level-1, level-3 is a best response to an opponent that plays level-2, and so on (Nagel, 1995). The variant of level-k which best matches the empirical similarity classes from the experimental data is level-$k(\alpha)$, which is an extension of level-1(α) from Fudenberg and Liang (2019). In practice, this makes level-$k(\alpha)$ a tie-breaking rule which selects the action with the lower variance when both actions have the same expected payoff.

[6] We define an outcome to be fair and efficient in the sense of near-equal split (NES) if, among the Pareto efficient outcomes, it minimizes the payoff differences between players.

$\mathbf{g_{58}}$	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 relative frequencies in bold. Underlined payoffs indicate Nash equilibrium outcomes

Dominance, which identifies four (DD, OD, CO, and MP).

One feature of our experimental design is that each subject plays all 78 games in our domain, including both sides of asymmetric games, with no feedback on other participants' play. By having subjects play both roles in each game, each participant makes a total of 144 distinct game choices. To the best of our knowledge, this represents the largest number of choices per subject in the behavioral game theory literature.[7] This design not only allows us to compute the empirical similarity classes discussed above, but also facilitates a detailed analysis of individual-level behavior, revealing systematic differences in how subjects perceive and respond to different games.

Consistent with the aggregate results, we find that, at the individual level, the majority of subjects' behavior is consistent with level-$k(\alpha)$ reasoning. However, subjects tend to deviate from level-$k(\alpha)$ reasoning when level-1(α) and Nash equilibrium actions yield outcomes that are neither fair nor efficient. In these cases, subjects who otherwise follow level-$k(\alpha)$ reasoning instead choose a near-equal split. This relationship between level-k and fairness also operates in reverse: subjects who typically favor playing near-equal split throughout the games tend to only deviate from that strategy to adhere to level-$k(\alpha)$.

While empirical similarity supports prediction transfer to nearby games (continuity or discontinuity of behavior), theoretical similarity indicates where a model predicts stability. Comparing them reveals which theoretical distinctions matter empirically. Our findings reveal important structural distinctions between level-k reasoning, rationalizability, 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 across games. While the level-$k(\alpha)$ model provides the overall best match for empirical similarity classes, incorporating efficiency and fairness considerations is necessary to account for certain deviations in behavior. Empirical similarity embodies two types of effects: the fact that small changes in payoffs may lead rules to take different actions, but also the fact that small changes in payoffs may lead subjects to adopt a different rule or heuristic.

Example. To illustrate our approach, consider the four coordination games shown in Figure 1.

[7] Given the number of choices subjects make, one might expect fatigue or learning effects that could render their choices inconsistent across the session. However, our experimental design accounts for this, and we find no evidence of systematic behavioral changes. We discuss this in Section 5 and later in Section C of the Appendix.

Each game admits two pure Nash equilibria: a Pareto-efficient equilibrium (A, A) and a Pareto-inferior equilibrium (B, B). As it turns out, in our neighborhood structure, 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}), while 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 level-$k(\alpha)$, all players choose to play an equilibrium for all k (both players play A in g_{60}, both players play B in g_{70}). Since level-$k(\alpha)$ groups games in which players converge to NE at $k = 1$ together (A1), g_{60} and g_{70} are theoretically similar according to level-$k(\alpha)$, while g_{60} and g_{70} are not empirically similar: while g_{60} is empirically similar to most A1 games which concentrate play on (A, A), g_{70} breaks away into the B2 class because the observed frequencies are roughly balanced between actions for both players. The efficiency motive appears empirically important yet is not accounted for by level-k. By contrast, g_{58} and g_{59} are both CO games in which level-$k(\alpha)$ leads to off-equilibrium outcomes (A3) in which both players can play A or B. Empirically though, they turn out to be similar to the A2 games instead of the A3 because the frequencies are concentrated over (A, A) and (A, B). Note also that, although g_{59} is a neighbor of both g_{58} and g_{60}, there is continuity in level-k rule and empirical behavior, respectively, only between g_{59} and g_{58}, not between g_{59} and g_{60}. Hence, they are in different rule based and 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.

Related literature. The literature in behavioral game theory has been largely motivated by the observation that human behavior in experiments frequently deviates from the standard Nash equilibrium prediction.[8,9] In many cases, the development of new behavioral rules and models have been driven by the selection and design of experiments which are specifically constructed to challenge whatever is the prevailing economic theory of strategic behavior. While this approach has significantly advanced our understanding of behavior, it raises 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 via meta-analysis. For instance, Wright and Leyton-Brown (2019) study behavior from normal-form games in the exper-

[8]Notable examples include ultimatum games (Gueth et al., 1982), beauty contest games (Nagel, 1995), 2×2 games with unique mixed equilibria (Erev and Roth, 1998), 3×3 games (Stahl and Wilson, 1994, 1995; Costa-Gomes et al., 2001), and the diverse experimental sets in Goeree and Holt (2001).

[9]Observed heterogeneity in human behavior has motivated various solution concepts, including 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). Learning models (Roth and Erev, 1995; Camerer and Ho, 1998) are omitted from this discussion due to the "initial play" nature of our experiment.

imental literature.[10] This is one of few papers that discuss the interaction between fairness and level-k, though they interpret playing NES as a feature of level-0 players which then affects how higher level-k players respond. We distinguish between NES and level-k as different behaviors, reserving the classification of level-0 only for behavior which is statistically indistinguishable from being random. The main challenge to meta-analyses is that, by construction, they only study games which researchers deemed sufficiently interesting to be studied. If most relevant strategic interactions are uninteresting to researchers, meta-analysis can limit our ability to identify frequently used behavioral rules.

Another approach is to study games with randomly generated payoffs, which ensures a broader coverage of games, but comes with its own limitations. Some notable examples include Zhu et al., (2025) with 2×2 games, Selten et al. (2002), and Fundeberg and Liang (2019) with 3×3 games, and Erev and Roth (1995) with constant-sum 3×3 games. The major drawback of randomly generated games is that they are likely to omit games which are important for identifying behavior but unlikely to be randomly drawn. For example, games with salient outcomes where both players can get the exact same payoff (e.g. the Prisoner's Dilemma, Stag Hunt) are often used by researchers to identify cooperative behavior, but they are unlikely to be drawn using random generation. [11] Fudenberg and Liang (2019) show that the modal behavior of their randomly generated games is best explained by level-1(α). It is only by randomly generating games in which level-1(α) is predicted to perform poorly that they find behavior in which subjects are more likely to induce efficient outcomes (Pareto Dominant Nash equilibria, or PDNE) instead of playing level-1.[12] One of our main contributions is that we use a complete domain of games, which allows us to mitigate the issues of both approaches: our selection of 2×2 games does not make omissions based on what has been studied before and it has a representative game for every possible strict preference ordering over outcomes, including many games which are highly unlikely to be drawn using random generation.

Similarity concepts have been studied before primarily in individual decision-making situations. Rubinstein (1988), Aizpurua et al. (1990), and Gilboa and Schmeidler (1995) theoretically study the similarity of judgments in individual decision-making. Evers et al. (2022) study similarity through categorization in mental accounting. A notable exception is Jehiel (2004), who introduces analogy-based expectation equilibrium for multi-stage games with perfect information.

Almost all experimental studies on similarity in games focus on a small selection of game classes. For example, Guida and Devetag (2013) study Prisoner's Dilemma- and Stag-Hunt-type games,

[10] Güth and Kocher (2014) review over three decades of variations of ultimatum bargaining experiments, highlighting fairness 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.

[11] Alon et al. (2025) further document biases and regularities, in particular in connection with dominance solvable games, when working with randomly drawn games. They show that there is no differences using their concepts whether games are randomly selected or also contain the omitted games.

[12] Camilo and Nagel (2026) show that NES is actually a better fit for the behavior in these games than PDNE: it not only significantly outperforms PDNE in predicting modal play, but it is also statistically comparable in performance to the bagged decision trees used as a benchmark in the original paper.

and Grimm and Mengel (2012) study acyclic 3×3 games with unique equilibria. The latter paper has similar results to our own, with level-$k(\alpha)$ and focal points determining which games are deemed similar. Ert et al. (2011) study the entire class of the simplest two-player extensive form games with binary choices for both players, drawing payoffs from $\{-8, -7, \ldots, 7, 8\}$ with all possible combinations of payoff structures. However, they do not provide a classification of the games based on aggregate behavior of subjects, or based on Nash equilibrium or level-1 reasoning.

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. Section 7 discusses extensions and further insights, and Section 8 concludes. Additional graphs, tables, proofs and formal definitions are provided in the Appendix.

2 Games and perspectives

Throughout the paper, we focus on 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 by relabeling players, relabeling 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 E 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 within a game, 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, by rewriting the bimatrix with the column player as the row player, the resulting payoff matrix is 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.[13] 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.

[13] 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 $\{1, 2, 3, 4\}$.

- **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 $\beta \in \Delta(A_{-i})$. [14]

Level-k rules and variants:

- **Level-0 (L0), Nagel (1995):** The mixed action $a_i^{L0} = (1/2, 1/2) \in \Delta(A_i)$ for player i, which randomizes uniformly over i's actions; a player who uses level-0 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.[15]
- **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 $0 < \alpha < 1$.[16] Notably, level-5(α) is equivalent to the NE action when it is unique and the L1(α) action otherwise in the set of games that we study.

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) [17]:** 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. [18]
  - **Self-favoring NES (sNES):** An NES where the acting player has a weakly greater payoff than the opponent.

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

[15] 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.

[16] 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.

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

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

- **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 $a \in A$.

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. In games with a unique NE, the NE is always PDNE. For coordination games with Pareto-ranked equilibria, the action leading to the Pareto optimal equilibrium outcome is chosen.

As a first step, we use these game-theoretical and behavioral rules to define theoretical and behavioral similarity classes of games, assuming that all subjects use a single rule or a 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 the 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 provides the feature sets (a la Tversky, 1977), according to which a subject attends to and chooses an action. Combining the metric space of games with a particular rule yields a subgrouping that partitions the set of games into similarity classes.

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 $e = \pi_2(a_{1,1}, a_{2,1}), \ldots, h = \pi_2(a_{1,2}, a_{2,2})$.[19]

Moreover, as mentioned before, there are a total of 7 more games represented by their respective 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. Thus, although they correspond to different tuples in $\mathbb{R}^8$, they all represent the same strategic interaction up to relabeling 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 (39) and (30), 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.

Let $\mathbf{\Gamma} = (\Gamma, E)$ 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 $\mathbf{\Gamma}$.[20]. 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.[21]

[19] 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.

[20] 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 $E' \subseteq E$

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

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 topology, which is studied in great detail in (44). 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'.[22]

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.

[22] Robinson and Goforth (44) 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.

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 $\psi \in \Psi$, $R(\psi g) = R(\psi g')$. [23]

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, 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 E 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 is 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, 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).

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

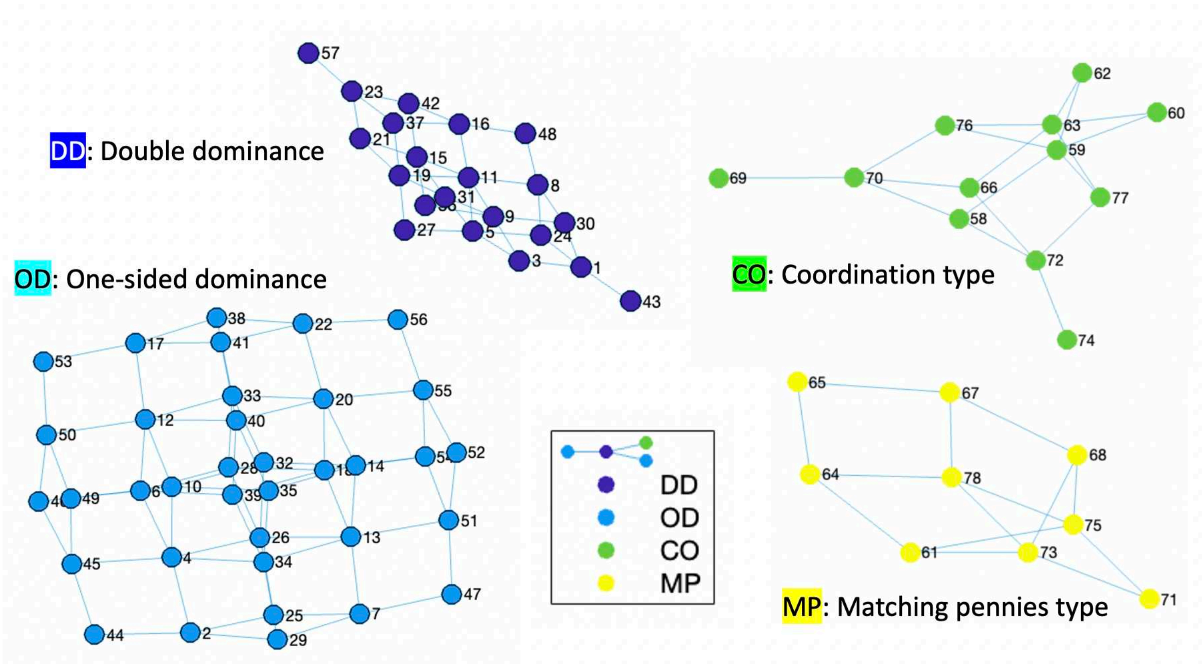

Figure 2: Graph of the similarity classes in G^* implied by the NSD rule. Each node is a game in G^*, and a link is drawn between a pair of games if: (i) for some representation of the pair of games in G, both games are neighbors, and (ii) both games belong to the same similarity class.

We refer to these four classes as the *strict dominance* similarity classes.[24] The reason CO games (green nodes) and MP games (yellow nodes) are separate, despite no actions being dominated in both classes, is due to the topology of neighboring games. Since there is no path from a CO game to an MP game that does not also pass through either DD or OD games (dark and light blue nodes) (see Figures A3 and A4 in the Appendix), the NSD rule has a discontinuity which makes CO games and MP games belong to different similarity classes. ■

The next similarity classes correspond to the Nash Equilibrium rule.

Nash Equilibrium similarity classes. If $R : G \to 2^A$ is a 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 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 support of a Nash equilibrium. The reason is the same as with the NSD rule, since it is

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

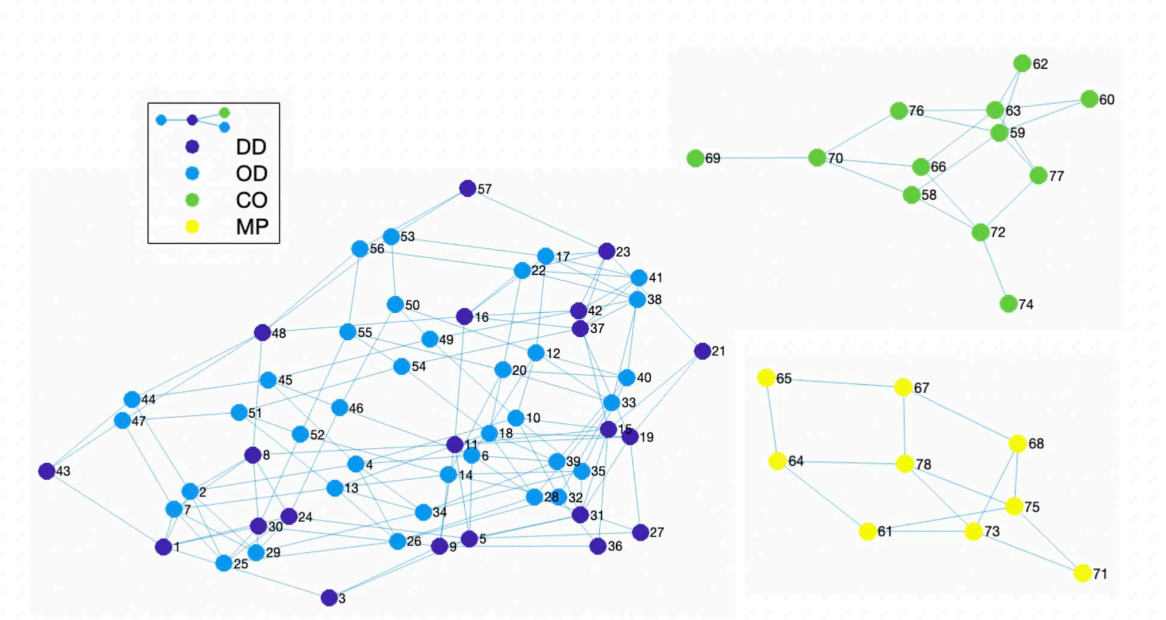

Figure 3: Graph of the similarity classes in G^* implied by Nash equilibrium. Each node is a game in G^* and a link is drawn between a pair of games if: (i) for some representation of the pair of games in G, both games are neighbors and (ii) both games belong to the same similarity class.

necessary to pass 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$k(\alpha)$, $k = 1, \ldots, 5$), we obtain different similarity classes from Nash. Taking all L$k(\alpha)$ rules for $k = 1, \ldots, 5$, as a set $\mathcal{R}$ of rules, leads to four similarity classes being distinguished (see Figure 4), namely:

- **One-outcome games (DD ∪ OD1 ∪ CO1) with** $(L1(\alpha), L1(\alpha)) = NE$**:** These are games where all the L$k(\alpha)$ actions of both players span only one outcome.[25] In particular, the L$k(\alpha)$ actions reach a pure Nash 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).[26]

- **Two-outcomes games (OD2) with** $(L1(\alpha), L2(\alpha)) =$ **NE:** These are games where L$k(\alpha)$ actions span two outcomes: player 1 always chooses the dominant action for any k and player 2's action converges to the unique Nash equilibrium at $k = 2$. These games consist only of OD games which we refer to as OD2 (component of light blue nodes in Figure 4; 18 games).

[25] 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, $\#\{a_1^{L1(\alpha)}, a_1^{L2(\alpha)}, \ldots\} \times \{a_2^{L1(\alpha)}, a_2^{L2(\alpha)}, \ldots\} = \ell$.

[26] 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^*.

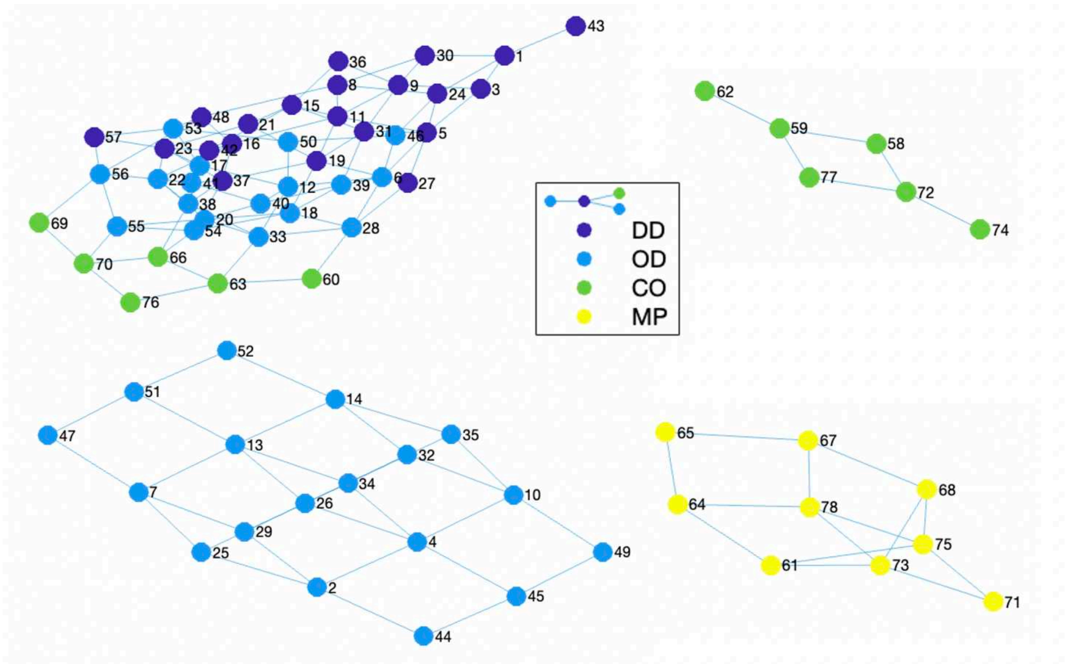

Figure 4: Graph of the similarity classes in G^* implied by level-$k(\alpha)$ rules, $k = 1, \ldots, 5$. Each node is a game in G^*, and a link is drawn between a pair of games if: (i) for some representation of the pair of games in G, both games are neighbors, and (ii) both games belong to the same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

- **Four-outcome games with two pure NE (CO2) where $Lk(\alpha) \notin$ NE $\forall k$:** These are games where L$k(\alpha)$ actions span all four outcomes and the Nash equilibrium is never reached by L$k(\alpha)$ at any step k. These games consist only of CO games which we refer 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 action profiles, spanning four outcomes. This coincides exactly with the MP games (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 classes is based on the number of outcomes a game yields when players play at 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 by a neighborhood-preserving map on G. Any path between a game in CO2 and a game in MP must pass through either the OD or DD component of 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 those obtained from Risk-Dominant Nash Equilibrium or from 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 and the implied empirical similarity classes of games. For this, define an **empirical rule with tolerance** δ, $\hat{R}_\delta : G \to 2^\Delta$, 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, which required exact equality in play behavior, we introduce a tolerance parameter, $\delta > 0$, to account 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 $(g_1, g_2, \ldots, g_n)$ 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 $\nu = 1, \ldots, n-1$, 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.

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

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) 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 in oTree (Chen, Schonger, and Wickens, 2016), and each subject participated in a standalone version. 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.[27] 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 players) across all 78 games (66 asymmetric and 12 symmetric). A screenshot of the decision interface is provided in Figure 5.

This design enabled us to reduce the number of required subjects while simplifying the selection and ordering of games for each participant. In a pretest, we evaluated whether playing all games (144 perspectives) rather than a smaller subset (44 perspectives) induced behavioral changes due to 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).

[27] 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.

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 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 clustering of theoretically similar games, which might lead subjects to expect future games to resemble previous ones, we introduced structure into the ordering of games rather than generating sequences purely at random. For example, if subjects encountered game g_1 and g_2 consecutively, both featuring the outcome $(4, 4)$ (see Section E 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, for example, toward 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" such that each letter corresponds to 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 E). 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. [28]

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, label randomization is performed as follows: for each perspective, we relabel 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)

[28] 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.

randomized sequences within subjects, we adopted this design to test whether the order within a sequence has a statistically significant effect.

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 rescaled to match the payoff scale of the games played in stage 1. The positive affine transformation adjusted the maximum lottery value in the task to correspond to the monetary value of receiving 4 points in the games, while the minimum value was adjusted to correspond to the monetary value of receiving 1 point. All values were presented in the currency of the subject's country to ensure consistency and comprehension.

5.3 Stage 3: Exit questionnaire

In the exit questionnaire, we asked participants to describe, in their own words, how they made decisions in the games. We also asked participants questions on a Likert scale about their perception of their strategy across stage 1 and collected information on their demographics (e.g. age, education, and gender identity).

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 $\times$ 2.67 average points).

6 Results

In this section, we document the results of our experiment. In Section 6.1 and 6.2, we argue that the theoretical similarity classes from Nash equilibrium and strict dominance fail to predict the discontinuities found by empirical similarity. We show that the empirical similarity classes we compute resemble the similarity classes implied by level-$k(\alpha)$. In particular, the level-$k(\alpha)$ classes are represented by four empirical similarity classes (A1-A4). We show that near-equal split helps explain the deviations from level-$k(\alpha)$ that we see in the remaining four empirical similarity classes (B1-B4). In Section 6.3 and 6.4, we show that aggregate and individual behavior is best explained

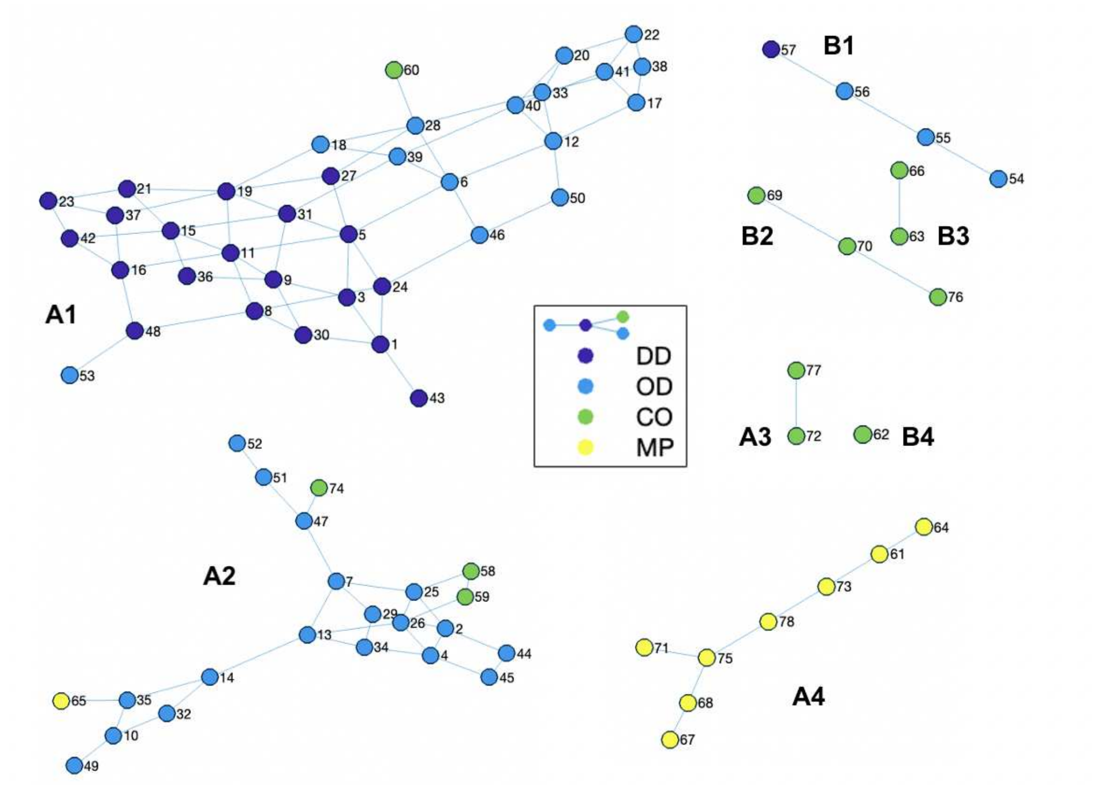

Figure 6: Graph of the empirical similarity classes in G^* from aggregate behavior of all subjects, for $\delta = 5\%$ at a 99.9% confidence level. Each node is a game in G^*, and a link is drawn between a pair of games if: (i) for some representation of the pair of games in G, both games are neighbors, 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.

by level-$k(\alpha)$ and near-equal split, and how heterogeneity both within and between subjects affect the empirical similarity classes.

6.1 Empirical similarity classes: a detailed description

As described in Section 4.5, estimating the empirical rule $\hat{R}_\delta$ (and thus, the resulting empirical similarity classes) requires selecting a threshold parameter δ. For the remainder of the paper, we set the threshold parameter δ to 5%. In the Appendix, we compute it for other threshold values, including 0%, 1%, and 2%. Figure 6 and 7 show the result, which we break down as follows:

A. Empirical similarity classes A1-A4: Level-$k(\alpha)$ related components

A1: This empirical similarity class consists of 36 games (20 DD, 15 OD1, and 1 CO1) from the one-outcome class implied by the Lk(α) similarity classes. With the exception of g_{17}, g_{37}, g_{48}, and g_{53}, level-$k(\alpha)$ converges to the unique Nash equilibrium at $k = 1$ and this equilibrium is also a near-equal split. In the remaining four games, the Nash equilibrium is Pareto efficient. The relative frequencies of subjects playing level-1(α) in this class of games is 67% or higher (top-right of Figure 7).

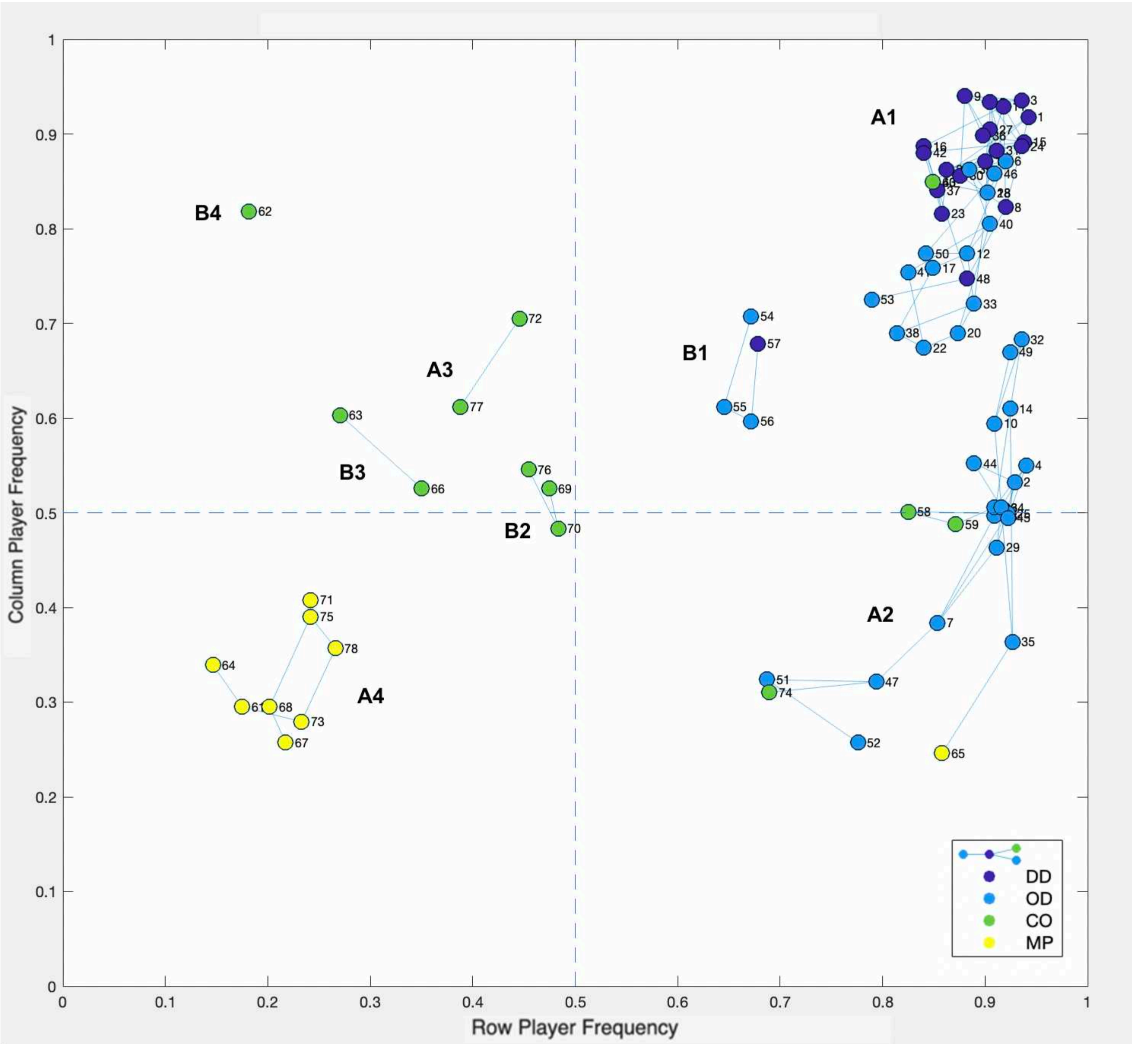

Figure 7: Empirical similarity classes with relative frequencies of play (renormalized for visibility). Each node is a game in G^*, and a link is drawn between a pair of games if: (i) for some representation of the pair of games in G, both games are neighbors, and (ii) both games belong to the same empirical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

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

A2: This empirical similarity class consists of 22 games (18 OD2, 3 CO2, and 1 MP), most of which are in the two-outcomes (OD2) class implied by the Lk(α) similarity classes. The relative frequencies of subjects are concentrated on level-1(α) for one player but split more equally across both actions for the other (center- and bottom-right of Figure 7). The empirical similarity of the CO2 (g_{58}, g_{59}, g_{74}) and MP (g_{65}) games to the OD2 games in this class can be explained by (i) the fact that g_{58}, g_{59} and g_{65} all have at least one perspective with a unique level-1 (without α) action, which concentrates play on a single action for at least one player; and (ii) the fact that g_{74} (Chicken) has an empirically salient near-equal split that is prescribed by level-1(α) that allows it to be empirically similar to the neighboring g_{47} in the OD2 class.

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

A3: This empirical similarity class consists of only 2 games (all OD2) from the four-outcomes class implied by the Lk(α) similarity classes, g_{72} and g_{77}. Both games have level-1(α) prescribe an action profile which is neither an equilibrium nor near-equal split. At the same time, these games have two salient near-equal split outcomes which gives an alternative explanation as to why the relative frequencies resemble the level-k(α) prediction for CO2 games (center of Figure 7).

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

A4: This empirical similarity class consists of all the MP games except for g_{65}. Unlike g_{65}, which has a unique level-1 action, these MP games all require level-1(α) to break at least one player's tie for selecting both actions. The relative frequencies of these games are concentrated around the action profile prescribed by level-1(α) (bottom-left of Figure 7).

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

While the level-k(α) similarity classes do not predict all the empirical components correctly, they do identify important regularities that are not predicted by other behavioral rules. For instance, level-k(α) correctly predicts that the OD1 and OD2 games are not similar; the same is true for the CO1 and CO2 games. By contrast, Nash and strict dominance would make no distinction between OD1 and OD2 games, or between CO1 and CO2. That being said, level-k(α), Nash, and strict dominance all correctly predict that MP games should be separated from other games. The remaining empirical similarity classes break away from the ones predicted by level-k(α) as follows:

B. Empirical similarity classes B1-B4: Breakaway components

B1: This empirical similarity class consists of 4 games (1 DD, 3 OD1) from the one-outcome class, most notably the Prisoner's Dilemma (PD) and other games in which the unique Nash equilibrium is not a near-equal split. The relative frequencies we find are comparable to other experimental studies of one-shot PDs, with about 30% of subjects colluding. These games break away from A1 because of the relative frequency increase in playing off-equilibrium actions (which induce near-equal split outcomes) from A1 to B1 (top-right of Figure 7).

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

B2: This empirical similarity class consists of 3 games from the one-outcome CO1 class that share similar properties to Stag Hunt. In these games, one of the equilibria is prescribed by all k of level-k(α) whereas the other equilibrium is near-equal split. The relative frequencies of this class are the closest to 50% of all the empirical similarity classes. These games break away from A1 because of the relative frequency decrease in playing level-1(α) actions from A1 to B2 (center of Figure 7).

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

B3: This empirical similarity class consists of 2 games from the one-outcome CO1 class that share similar properties to Battle of the Sexes. In these games, level-$k(\alpha)$ converges to an equilibrium at $k = 1$ that is near-equal-split but favors one player over another; at the same time, the other equilibrium is a near-equal split favored toward the opposite player. The relative frequencies of this class are weighed slightly toward playing the level-1(α) equilibrium. These games break away from A1 because of the relative frequency decrease in playing level-1(α) from A1 to B3 (center of Figure 7).

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

B4: This empirical similarity class consists of one game from the four-outcomes CO2 class that shares similar properties to Battle of the Sexes. In these games, each equilibrium is a near-equal split but level-1(α) prescribes selecting the equilibrium which is self-favoring; if both players play level-1(α), the action profile fails to coordinate on either equilibria. The relative frequencies of this class concentrate on the off-equilibrium level-1(α) prediction. These games break away from A3 because of the relative frequency increase in playing level-1(α) from A3 to B4 (top-left of Figure 7).

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

6.2 Comparison of empirical and key theoretical similarity classes

For the purposes of exposition, we describe the first four empirical similarity classes (A1-A4) as being related to level-k(α) due to the visual resemblance between Figure 4 and the A1-A4 classes in Figure 6. To properly quantify the extent to which level-$k(\alpha)$ predicts the empirical similarity classes, we use the adjusted Rand Index (ARI) from Hubert and Arabie (1984). The ARI is a widely used metric in computer science and network analysis for evaluating classification from clustering algorithms. In short, the ARI is a measure of similarity between two partitions, adjusting for the possibility of matching by chance. For every rule in our analysis, we calculate the ARI between the similarity classes of a particular rule and the empirical similarity classes.[29] The resulting indices are reported in Table 1.

Table 1: 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.

The ARI between the empirical similarity classes and the level-$k(\alpha)$ similarity classes is the highest among all behavioral rules, significantly outperforming the rest at the 1% level with the exception of level-k. The fit is more visually apparent when we compare how the games are distributed among different similarity classes empirically versus theoretically (Figure 8).

For example, consider the following simple classification fit statistic: for a rule R, let $S_R = \left(\sum_c \#S_c\right)/78$ where S_c is the largest set of empirically similar games contained within a theoretical similarity class c. This statistic can be visually calculated in Figure 8 by adding the maximal overlaps between the number of games in each empirical similarity class and the number of games within a theoretical similarity class. S_R for NE and NSD is 59% and 63% respectively, on account of correctly predicting the discontinuity in the MP games as well as the partial overlap between DD and OD games. However, both rules fail to correctly separate the OD games into

[29]We omit Hubert and Arabie's formula for the sake of brevity.

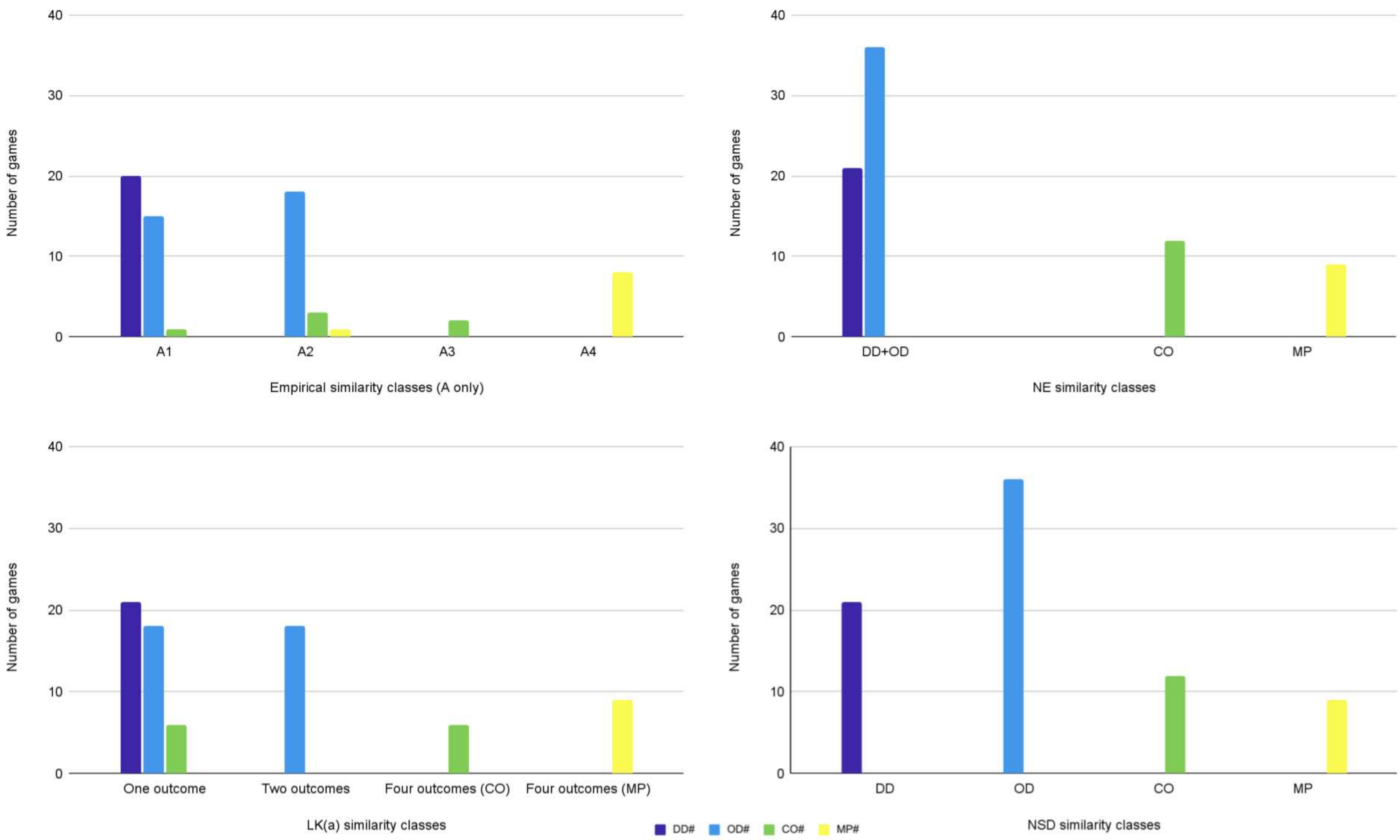

Figure 8: Distribution of games by similarity classes according to different rules, colored by strict dominance classes. The top-left are the first four empirical similarity classes that correspond with level-k(α), the top-right are the theoretical similarity classes according to Nash equilibrium, bottom-left according to level-k(α), and bottom-right according to strict dominance (reported as NSD).

OD1 and OD2 (or the CO games into CO1 and CO2). By contrast, level-$k(\alpha)$ correctly makes these distinctions and has an S_R of 82%.

That being said, the failure for level-$k(\alpha)$ to predict the breakaway classes (B1-B4) is, we argue, related to its tension with near-equal split. In each of the breakaway classes, the separation corresponds with either a substitution of level-k(α) play with near-equal split (e.g. A1 to B1, A1 to B3), the opposite (A2 to B2), or both simultaneously (A2 to B4). This is despite the fact that near-equal split is a poor predictor of the empirical similarity classes (ARI 0.03, 99% CI: -0.08, 0.16), significantly worse than both NE (ARI 0.36, 99% CI: 0.20, 0.55) and NSD (ARI 0.32, 99% CI: 0.18, 0.55). This raises important questions not only about the relationship between the two behavioral rules but also about the potential complementarity in how our subjects use them. We explore this further in the next section, in which we characterize the behavior we observe in our data, both at the aggregate and the individual level.

6.3 Other measures of aggregate behavior

The empirical similarity classes show how discontinuities in choice data from our experiment group the games into distinct classes. To connect these findings to the broader experimental literature, we complement our approach with standard methods for studying aggregate behavior.

One approach is to estimate what is the best performing rule relative to some measure of accuracy/error. Fudenberg and Liang (2019) (FL) use modal play as their measure of accuracy: for a rule R, they define the accuracy of rule R as the number of perspectives in which R out-of-sample predicts the most frequently played action.[30] If we use modal play in our data, we find that level-1(α) achieves the highest accuracy, correctly predicting 94.44% (136/144) of the most frequently played actions. The difference between level-1(α)'s accuracy and that of other rules in our analysis is statistically significant at the 1% level, with the exception of level-5(α) which is second-best (see Figure 9). Many experimental studies have found level-1 as a best predictor, though typically with much less support for higher level-k compared to our data (e.g. FL themselves find the same result in 3×3 games, Selten et al. (2003) with randomly chosen 3×3 games with the strategy method, and Ert et al. (2011) in simple extensive form games). When level-1(α) fails to predict modal play, both sNES and PDNE achieve perfect accuracy (see Table 2), which is also consistent with FL's result that when level-1(α) predicts poorly in randomly generated 3×3 games, the best predictor is PDNE.[31] Nonetheless, although level-1(α) accurately predicts modal play more than 90% of the time, its similarity classes are a poor fit (ARI 0.00, 99% CI: 0.00, 0.00) with the empirical similarity classes.

One of the reasons for the discrepancy between what predicts modal play (level-1(α)) and what predicts the empirical similarity classes (level-k(α)) is the fact that the latter relies on estimating

[30] Because none of our rules (except for QRE) involve estimating parameters that fit the data, simply calculating in-sample the number of perspectives in which each rule predicts the most frequently played action should allow us to infer which rule best predicts our data. For a discussion on in-sample vs. out-of-sample fit, see Mullainathan and Spiess (2017).

[31] Camilo and Nagel (2026) show that NES actually outperforms PDNE in this set of games.

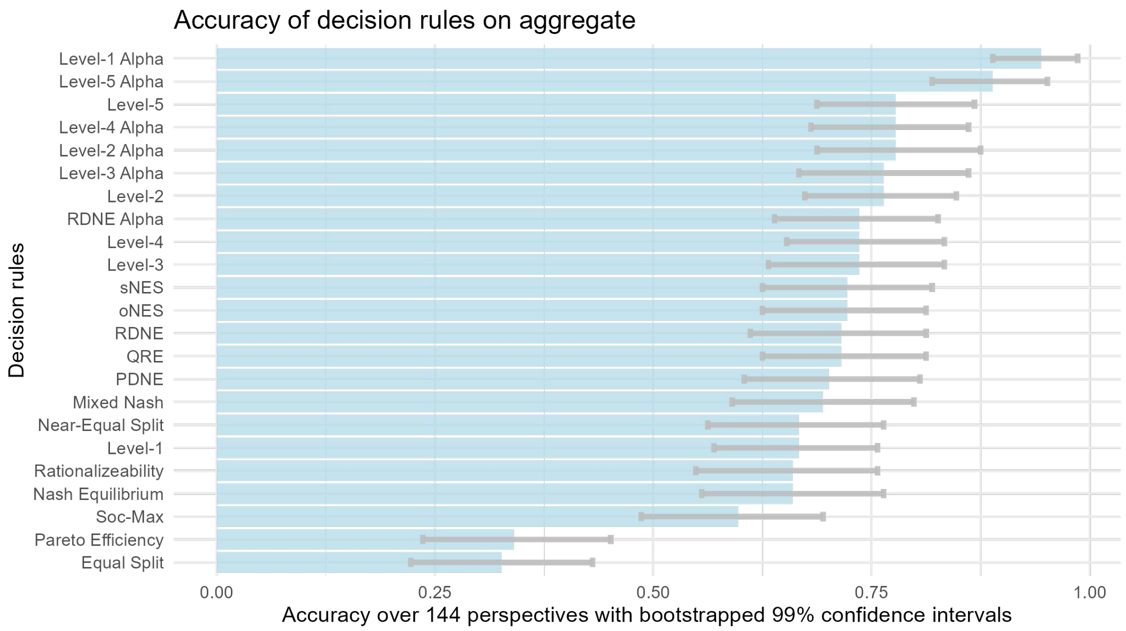

Figure 9: Accuracy of decision rules in predicting the most frequently played action.

the frequency distribution whereas the former relies only on the coarser statistic of which action has more weight. With this in mind, we repeat the same exercise but using a mean-squared error (MSE) measure instead of modal play accuracy: for a rule R, we average the squared difference between R and the observed frequency distribution across all perspectives.[32] Using MSE, we find that the lowest MSE rule is level-1 (see Figure 10). Unlike level-1(α), level-1's similarity classes have some fit with the empirical similarity classes (ARI 0.37, 99% CI: 0.23, 0.61). Among the rules that predict only one action per perspective, level-1 has the highest ARI though its difference in ARI to that of most other rules is not statistically significant. However, level-1's ARI is still lower than the ARI of level-k(α) and even the ARI of level-k. The fact that level-k significantly outperforms level-1 as well as all other rules that predict only a single action per perspective might suggest that either subjects vary in the rules that they adopt or that subjects might be using a combination of rules simultaneously. Evaluating either hypothesis requires characterizing the behavior of our subjects at the individual level.

6.4 Individual behavior and behavioral rules

To analyze the behavior of each of our 450 subjects across the 144 choices they make in every perspective, we aim to classify each subject according to the behavioral rule that best fits them. To infer whether subjects use a combination of rules, we examine what rules subjects use when they deviate from that which best fits them. Finally, we also analyze whether our subjects play in a way which is significantly different from uniform random play (level-0) so as to limit the chances

[32]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.

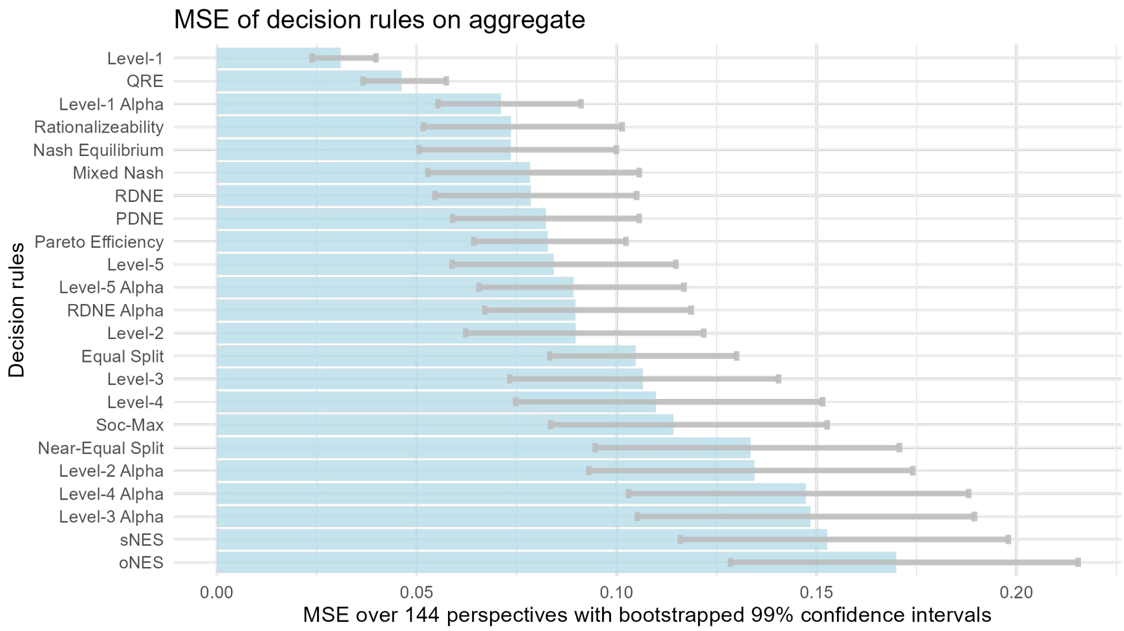

Figure 10: MSE of decision rules in predicting the frequency distribution of perspectives.

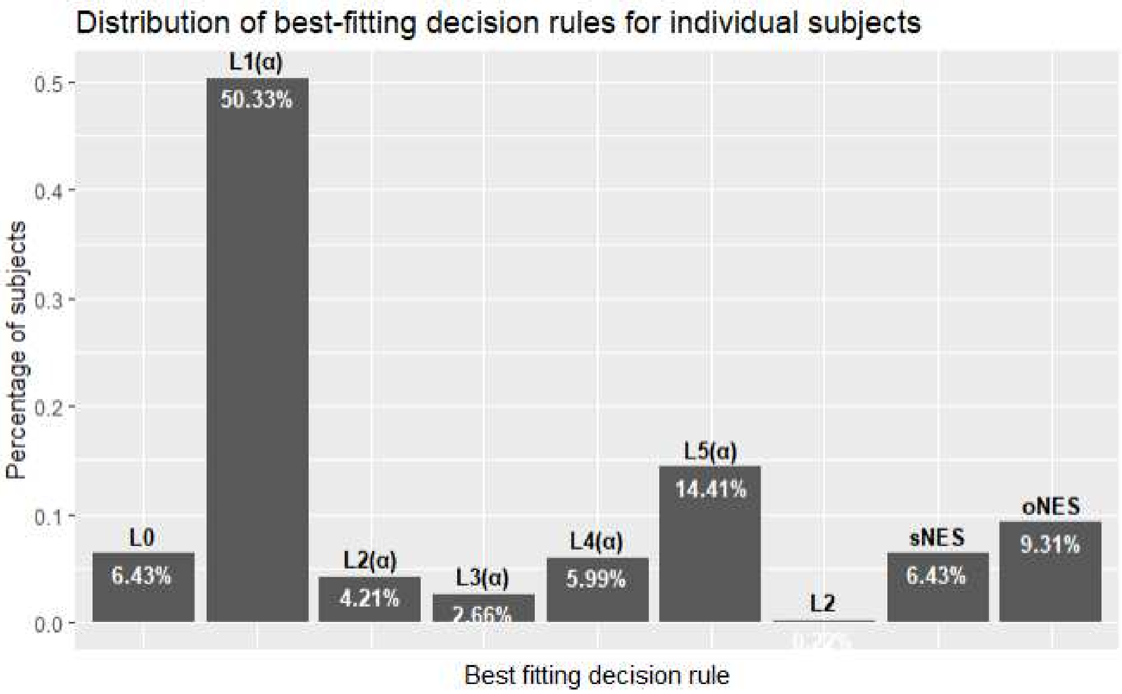

Figure 11: The distribution of all 450 subjects by the best-fitting rule. Note that L5(α) is a rule identical to choosing NE action when it is unique and L1(α) otherwise.

Table 2: 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	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

of false classifications.[33]

The distribution of subjects' best-fitting decision rules is summarized in Figure 11. Level-1(α) emerges as the most common best-fitting rule among individual subjects, followed by the two near-equal split 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 12. 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 appear to combine these two rules to varying degrees.[34]

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 12).

The complementarity between level-$k(\alpha)$ and near-equal split provides the framework for under-

[33] We classify a subject as level-0 if we fail to reject the hypothesis that the best-fitting rule has an accuracy greater than 50% at a 95% confidence level. The confidence level is reduced due to the lower number of observations per subject.

[34] 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.

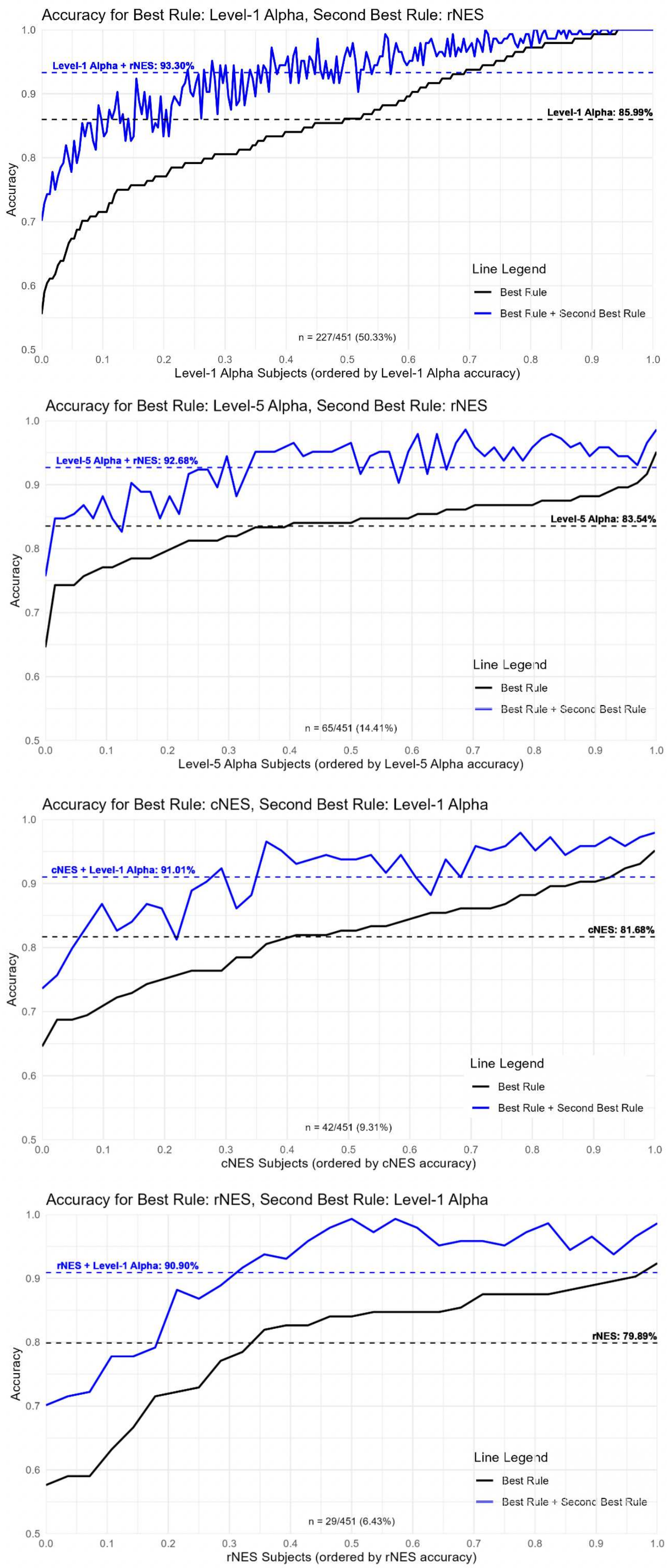

Figure 12: The quantile function of the accuracy of each subject's best-fitting rule (black) plus the added accuracy of other rules (blue) in games where the best-fitting rule predicts incorrectly. This figure reports sNES and oNES as rNES and cNES respectively.

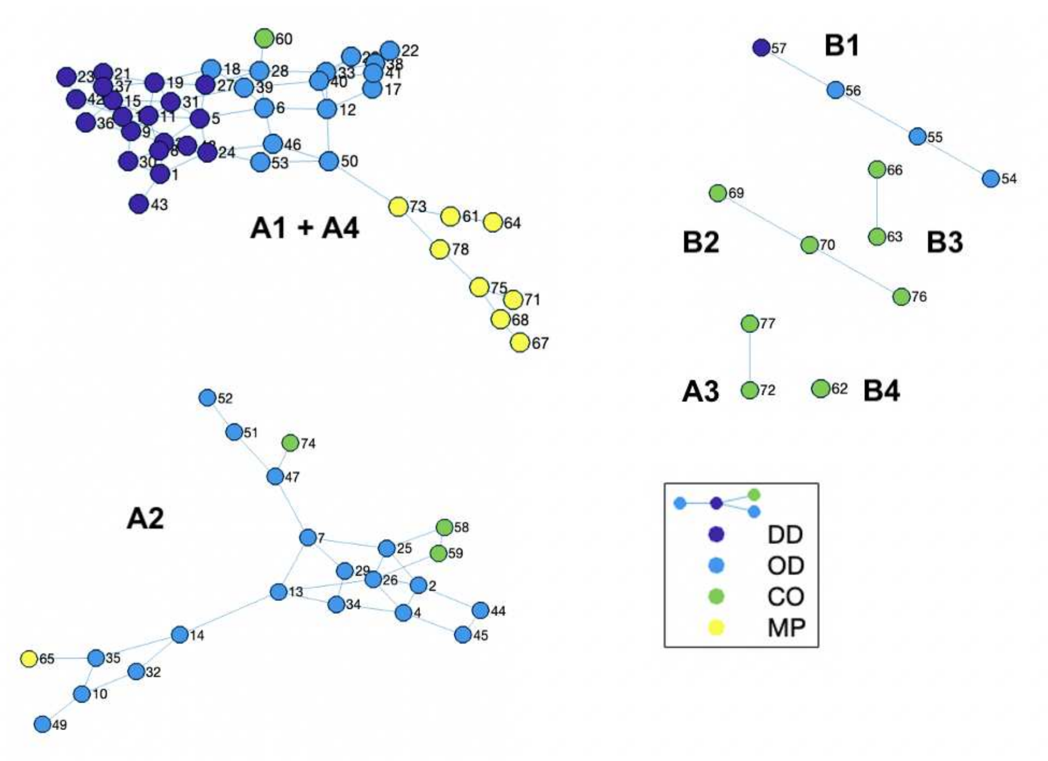

Figure 13: Empirical similarity classes in G^* based on aggregate behavior of subjects primarily identified as level-$k(\alpha)$ types, for $k = 2, \ldots, 5$, and for $(\delta, p) = (5\%, 99.9\%)$.

standing why the empirical similarity classes further refine the similarity classes implied by level-$k(\alpha)$. For instance, the fact that the breakaway classes separate from A1-A4 due to changes in the overlap between near-equal split and level-k(α) can now be explained by not only the 15% of subjects best fit by near-equal split rules, but also the other subjects who frequently deviate from their best fitting rule to play near-equal split.

6.5 Individual behavior and similarity classes: A first application

In Section 6.4 we showed that there is heterogeneity between subjects and also that individual subjects combine different rules. By applying empirical similarity to groups of players that share a best fitting rule, we are able to provide insight as to whether the discontinuities found on the aggregate are a result of subjects combining rules as opposed to subject heterogeneity.

More concretely, if it were the case that the discontinuities on the aggregate were purely driven by subject heterogeneity, we would find that the empirical similarity classes for players best fit by a rule R should largely resemble the similarity classes predicted by R. On the other hand, if the empirical similarity classes for players best fit by R resemble the empirical similarity classes from the full sample and, moreover, this is also the case for subjects fit by other rules, this would suggest that the discontinuities are driven by subjects combining rules.

We compute the empirical similarity classes for different subgroups of subjects. The empirical similarity classes for subjects best fit by level-1(α) are identical to the similarity classes predicted by level-1(α), which consist of all games belonging to the same similarity class (see Figure A23 in the Appendix). This means that within this group of subjects, we do not detect statistically significant deviations that produce the discontinuities we found in the breakaway classes. By contrast, the empirical similarity classes for subjects best fit by level-$k(\alpha)$ for $k > 1$ are essentially the same as the empirical similarity classes with all the subjects, except for one extra link between the A1 and A4 games. These subjects follow level-$k(\alpha)$ on a large subset of games but deviate from those rules precisely on the breakaway games (B1-B4). This similarly applies if we consider all level-$k(\alpha)$ players, wherein the empirical similarity classes are comparable to those found on aggregate, but with more deviations than those found among $k > 1$ classified subjects (Figure 13).

7 Discussion

Our notion of similarity presents a framework for grouping games based on the continuity of behavior, which we apply to a comprehensive domain of 2×2 games. We studied how different models from game theory and behavioral economics predict different similarity classes, and we showed how those models' similarity classes compare to that estimated from our own experimental data. By having subjects play all of the games within our domain as opposed to playing either games selected from the literature or randomly generated games, we were able to identify individual differences between subjects from which to document the diversity and structure in how people generally play games. That being said, we foresee several applications, extensions, and refinements that can also be explored using our framework.

7.1 Insights from similarity classes

Generalizability of canonical games. By definition, our notion of similarity requires similar games to have continuity in behavior. This directly allows us to quantify the extent to which behavior generalizes and remains predictable across games within our domain. For example, consider the "canonical games" which are widely used both in experiments and in the classroom: the Prisoner's Dilemma (g_{57}), Stag Hunt (g_{69}), Battle of the Sexes (g_{62}), and Matching Pennies (g_{67}). Each of these games belong to small empirical similarity classes of their own: g_{57} is in B1 ($4/78 \approx 5.1\%$), g_{69} in B2 ($3/78 \approx 3.8\%$), g_{62} in B4 ($1/78 \approx 1.3\%$), and g_{67} in A4 ($8/78 \approx 10.3\%$).[35] If we assume that our domain of 2×2 games is representative of the strategic interactions people actually encounter in the real world, the fact that these empirical similarity classes are small suggest that the behavior in the canonical games are not broadly representative of how people play games. This limits the flexibility of games we can use to induce behavior that is similar to what we find in the canonical games. At the same time, this also means that the larger similarity classes warrant special attention, as they showcase a wide variety of 2×2 games that exhibit the desirable property that play concentrates reliably on either one or two action profiles.

[35] There are other games comparable to Battle of the Sexes and Matching Pennies, as we described earlier in Section 6.1. Considering the alternative representations does not change the broader point that these games belong to small empirical similarity classes.

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 different behavioral rules. 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 that clearly separate them may yield deeper insights into how these rules complement each other.

Mechanism design. Brandenburger and Nalebuff (1995) argue that strategic success often comes from changing the game rather than merely playing it. We expect that one of the main applications of our empirical similarity classes will be as a practical tool for mechanism design. For example, using our experimental data and the neighborhood structure alone, a mechanism designer can find small targeted changes to induce a particular outcome $\tilde{\pi}$. In principle, the designer can (i) find a game in our dataset that best represents the strategic interaction of interest, say g_0 with outcome π_0, and then (ii) select the game $\tilde{g}$ which most closely induces $\tilde{\pi}$. The neighborhood structure directly provides the designer with the minimal sequence of payoff changes necessary by finding the shortest path from g_0 to $\tilde{g}$. Similarity provides the added benefit of informing the designer which games on the path are likely to lead to behavioral changes which may be undesirable. We provide a demonstration of how this example can be implemented in Section **??** of the Appendix.

One of the advantages of this approach is that it implements outcomes directly based on the observed behavior, as opposed to relying primarily on changing either the equilibria or the strictly dominant action.

7.2 Robustness checks

Payoff structures. Our analysis only studies games with payoffs drawn from {1,2,3,4} without replacement, which our subjects treat as cardinal payoffs. The robustness of our results to different payoff structures is an important but open question. Zhu et al. (2025) (Z25) study a slightly different set of 2×2 games: in their setting, payoffs are drawn from [0,50] but with the same preference orderings represented as in Robinson and Goforth (2005) (RG5), excluding the MP games with the caveats mentioned in the introduction. Since our games are the same as RG5's, this allows us to compare Z25's data to our own to partially address this question.

We use all the games from Z25's dataset that have (i) the same preference orderings as g_{58}, g_{59}, g_{60} and g_{70}, and (ii) the same level-k (and level-k(α), $\alpha < 1$ predictions as we do for those games. These are the four coordination games we discussed in the Introduction. Our prediction is that games with the structure of g_{60} should have relative frequencies in the one-outcome quadrant $[0.75, 1] \times [0.75, 1]$ (like A1), games with the structure of g_{58} and g_{59} should be in the two-outcome

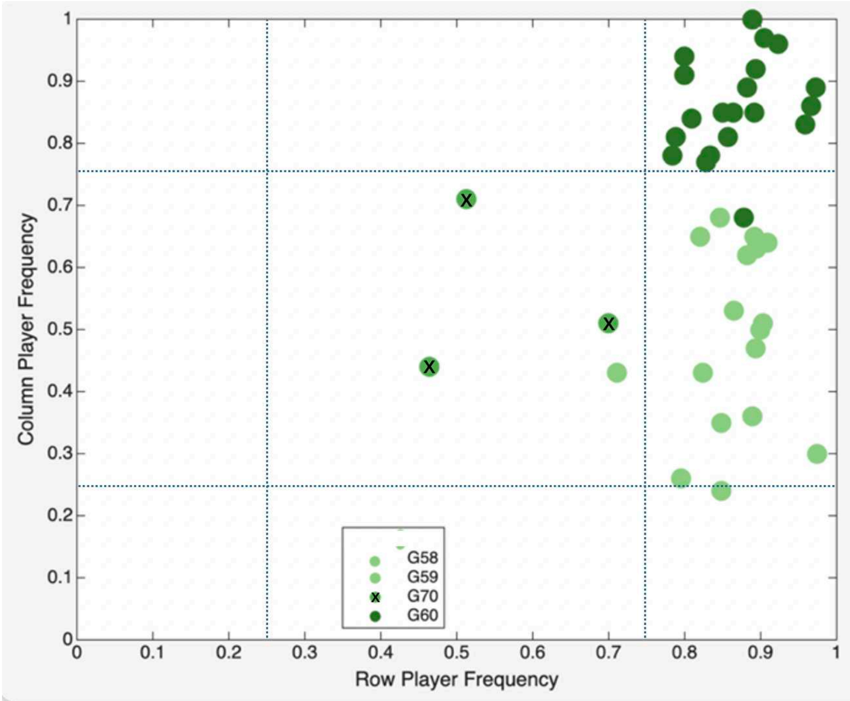

Figure 14: Frequencies of play (renormalized for visibility) in games taken from Zhu et al. (2025) with payoffs in [0, 50] and payoff structures corresponding to the ones of games g_{58}, g_{59}, g_{60} and g_{70} respectively.

quadrant $[0.75, 1] \times [0.25, 0.75]$ (like A2) and games with the structure of g_{70} should be in the four-outcome quadrant $[0.25, 0.75] \times [0.25, 0.75]$ (like B2) (see Figure 7). We find that of the 40 games from Z25 that we study, 38 of them fit our prediction: 19/20 for the A1-like games, 15/16 for A2, and 3/3 for the B2 (see Figure 14).

Expanding the behavioral rule set While our analysis includes a broad set of behavioral rules, there are omissions that warrant further study. For instance, while our analysis suggests that the standard QRE model from McKelvey and Palfrey (1995) does not explain the empirical similarity classes well (see Figure A16), there are other variants of QRE that could. Incorporating extensions of QRE may better explain behavior in certain games, especially those with mixed-strategy equilibria. At the same itme, incorporating hybrid rules that combine multiple models may account for the observed within- and between-subject variation. Developing these extensions opens the door to designing new experiments to test these models and refining our methodology to accomodate 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 we develop in this paper can, in theory, be applied to other game domains, including 3×3 games. In practice, however, there are computational challenges that arise as the number of games within a domain 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 in the same way that we do in this paper.

This limitation can nonetheless be partially overcome by considering a parametric family of games. For instance, consider the set of two-player linear best response games with the following payoff function:

$$\pi_i = c_i + \beta_i a_i + \theta_i a_i^2 + \gamma_i a_i a_j,$$

where $B_i = (c_i, \beta_i, \theta_i, \gamma_i) \in \mathbb{R}^4$ is the parameter vector for player $i = 1, 2$ and $a_i \in [0, 1]$ is player i's action. This parametric family includes many games of interest, including the generalized Beauty Contest and models of Cournot (Mauersberger and Nagel, 2018). By restricting the parameters to a small finite subset of $[-1, 1]$, we can define a topology based on the distance between the parameter values across different games to study how our framework applies to this domain. Provided that the topology and the set of games are well-defined, our notion of similarity, including its ability to produce similarity classes, can be readily adapted to different domains.

8 Conclusion

This paper makes several novel contributions to the study of strategic interaction. First, it examines human behavior across a comprehensive domain of one-shot 2×2 games, with each participant deciding on every game in the set. This makes it one of the largest experiments of its kind on one-shot games. Second, it employs a geometric approach to group games into theoretical and empirical similarity classes. These classes capture both the predictions of behavioral and game-theoretic solution concepts and the observed behavior of participants.

Third, we systematically integrate two pivotal behavioral rules— level-$k(\alpha)$ and near-equal-split — which are typically studied in isolation. The complete domain of 2×2 games allows us to observe the interaction of these distinct behavioral rules in sufficiently rich environments to distinguish between strategic reasoning, fairness-based motives, or a combination of both.

We find that empirical similarity classes derived from aggregate data align closely with theoretical classes based on level-$k(\alpha)$ rules. Yet, near-equal split plays an important complementary role, particularly when NES outcomes diverge from level-$k(\alpha)$ or Nash outcomes. The capacity of level-$k(\alpha)$ rules to explain individual behavior is demonstrated across a broader range of games than previously documented, supporting their applicability across diverse strategic contexts. Given the large number of games observed per subject, we identify behavior that adheres to multiple behavioral rules, moving beyond the "single-rule plus noise" paradigm prevalent in existing literature.

Our analysis also clarifies the predictive roles of different solution concepts. The shortcomings of Nash equilibrium and strict dominance compared to level-$k(\alpha)$ are apparent, particularly in the classification of dominance-solvable games. Whereas strict dominance separates one-sided from two-sided dominance games, Nash equilibrium fails to distinguish them. The empirical similarity classes suggest that a more salient distinction lies between games solvable in one or two steps of level-$k(\alpha)$ reasoning (DD and OD1) versus OD2 games. Matching pennies games, by contrast, are correctly classified by all concepts into one separate empirical similarity class.

As noted in the experimental literature, coordination games remain among the most challenging to predict. Level-$k(\alpha)$ distinguishes two classes of coordination games based on whether the NE is reached in one step or more. Yet, efficiency and fairness considerations appear to play a role in coordination that is not captured by strict dominance, Nash, or level-$k(\alpha)$ rules.

Much like a geographical atlas provides different thematic maps—topographical, political, or climatic—of the same territory, our behavioral atlas offers multiple structural perspectives on the same set of games. By layering behavioral rules (level-k, NES), game-theoretic benchmarks (Nash Equilibrium, strict dominance, Pareto optimality), and empirical data, we create a multi-dimensional map of strategic interaction. Comparing these layers allows us to identify the specific strengths and weaknesses of each framework and, crucially, to pinpoint which subjects drive observed empirical regularities.

When we deconstruct the aggregate similarity classes by primary behavioral rule, a striking pattern emerges. The $L1(\alpha)$ subjects—the largest group in our sample—produce a similarity structure consisting of only a single, undifferentiated class. By contrast, we find that the $L_{k>1}(\alpha)$ subjects are primarily responsible for the breakaway games observed in the aggregate data. Their specific similarity structure maps most closely to the overall empirical results, demonstrating that while $L1$ types may be more numerous, the sophisticated $L_{k>1}$ types are the architects of the empirical diversity we observe across game classes.

This "behavioral atlas" provides a starting point for researchers and practitioners interested in predicting, comparing, or modeling human decisions. By systematically mapping empirical regularities, this study offers both a benchmark for future theoretical work and a foundation for extending behavioral models to broader domains. It also facilitates game design applications. Following Brandenburger and Nalebuff's (1995) insight that strategic success often stems from changing the game rather than merely playing it, our atlas enables the identification of "neighboring games" to 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 informed local search in payoff space.

Looking ahead, future research should test the centrality of Level-$k(\alpha)$ and NES rules in richer contexts, investigate the stability of empirical similarity classes beyond 2×2 games, and explore the multiplicity of rules 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 improve predictions of human decision-making. For such a program, AI and machine learning will be indispensable, but an anchoring in the fundamental knowledge of human behavior remains essential.

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] **Cerigioni, Francesco.** 2021. "Dual Decision Processes: Retrieving Preferences When Some Choices Are Automatic." *Journal of Political Economy.* 129(6): 1667-1704.

[14] **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.

[15] **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.

[16] **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

[17] **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.

[18] **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.

[19] **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.

[20] **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.

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

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

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

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

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

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

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

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

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

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

[31] **Hubert, Lawrence and Phipps Arabie.** 1985. "Comparing partitions.," in *Journal of Classification.* 2: 193-218.

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

[33] **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.

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

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

[36] **Mullainathan, Sendhil and Jann Spiess.** 2017. "Machine Learning: An Applied Econometric Approach," *Journal of Economic Perspectives*, 31(2): 87-106.

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

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

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

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

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

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

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

[44] **Robinson, David and David Goforth.** 2005. *The Topology of* $2{\times}2$ *Games: A New Periodic Table.* Routledge.

[45] **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.

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

[47] **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.

[48] **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.

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

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

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

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

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

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

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

[56] **Wright, Kevin and Kevin Leyton-Brown.** 2019. "Level-0 meta-models for predicting human behavior in games," *Journal of Artificial Intelligence Research 64,* 357–83.

[57] **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 $\psi g'$ in $(G, E(A_N \odot A_R))$ such that the rule is constant for some $\psi \in \Psi$. At the same time, note that $\psi g'$ is in the same component of $(G, E(A_\Psi))$ as g'. Since $E(A_N \odot A_R), E(A_\Psi) \subset E(A_\Psi + (1 - A_\Psi) \odot A_N \odot A_R)$, this admits a path from g to g' in $(G, E(A_\Psi + (1 - A_\Psi) \odot A_N \odot A_R))$ (namely, using first the path through $E(A_N \odot A_R)$ from g to $\psi g'$ and then going from $\psi g'$ to g' through $E(A_\Psi)$). 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

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

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 conditions 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 $g \sim_R g'$. □

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).

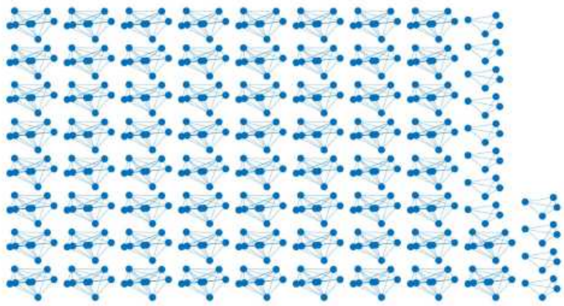

Figure A1: The graph on G, where edges represent equivalence between two games. Each of the 78 components consists of 4 or 8 equivalent games in G, depending on whether the component is one of symmetric or asymmetric games. Formally, this graph is denoted as $(G, E(A_\Psi))$.

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 is 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

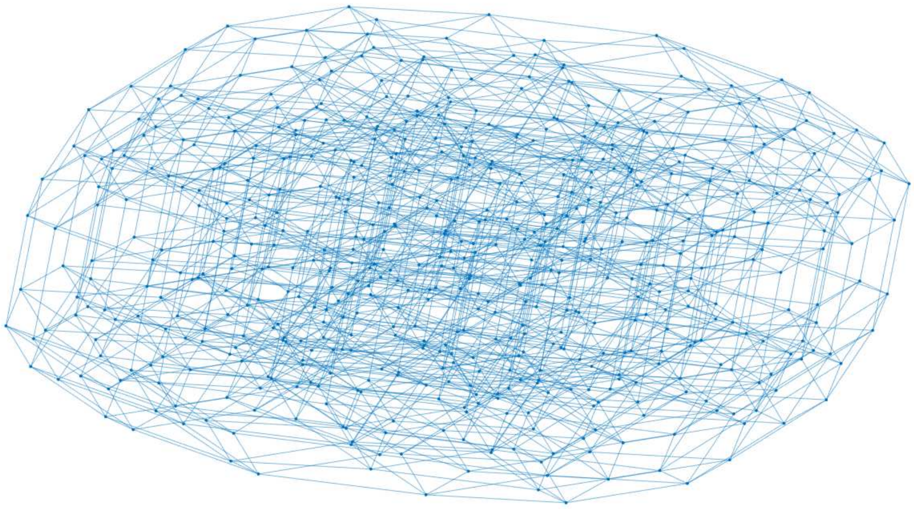

Figure A2: The graph on G, where edges represent whether two games are neighbors in the Robinson-Goforth topology. Formally, this graph is denoted as $(G, E(A_N))$.

outcome, always yield a unique similarity class, comprising all games in G (and similarly for G^*). However, combining the two rules yields nine similarity classes; see Figure A13. The same nine classes are obtained 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) consists of 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;

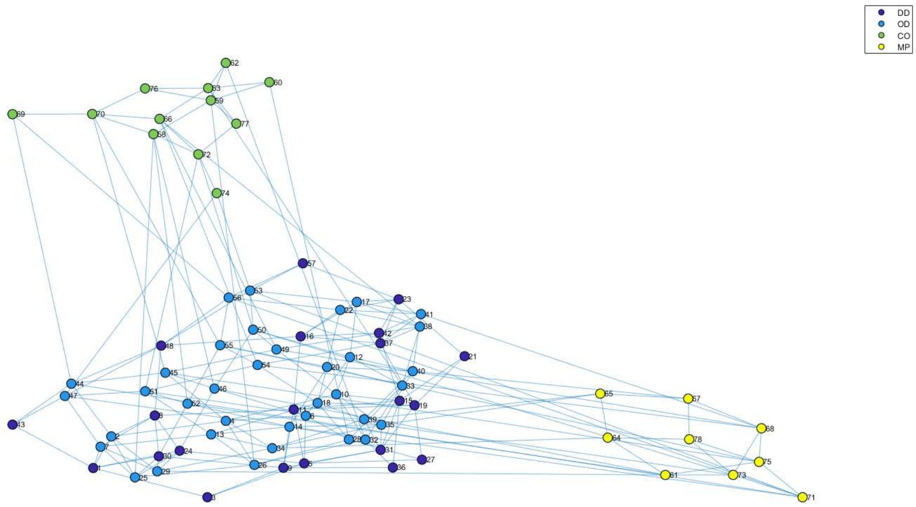

Figure A3: Graph of the Robinson-Goforth topology in G^* colored by types of games, with nodes representing the similarity classes according to Nash equilibrium and edges showing how they connect. Here, we also include edges between games across the three game classes. For ease of reading, they are not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

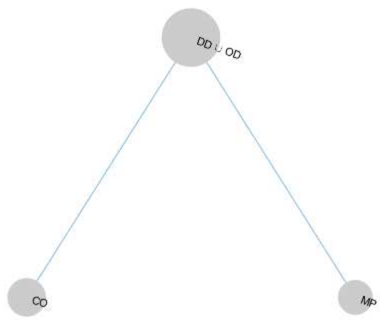

Figure A4: Graph of the Robinson-Goforth topology in G^*, with nodes representing the similarity classes according to Nash equilibrium and edges how they connect.

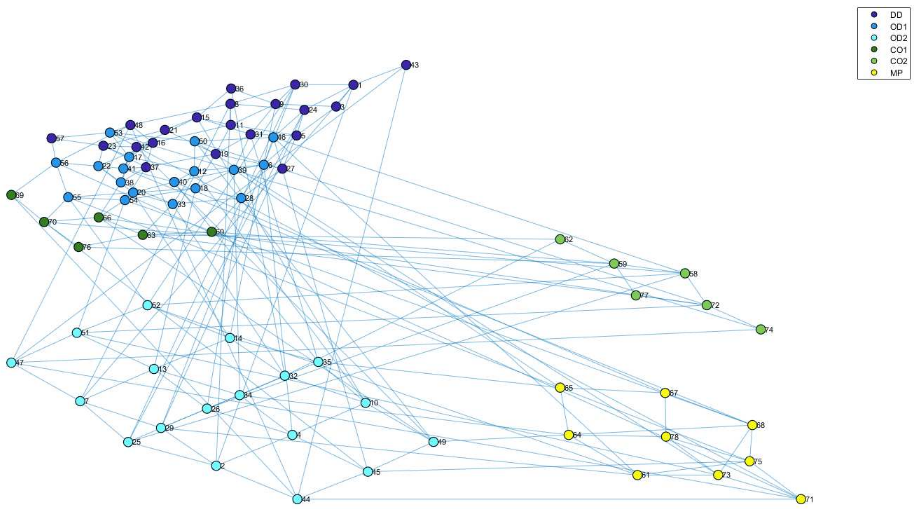

Figure A5: Graph of the Robinson-Goforth topology in G^* colored by similarity classes implied by Level-$k(\alpha)$ rules, $k = 1, \ldots, 5$.

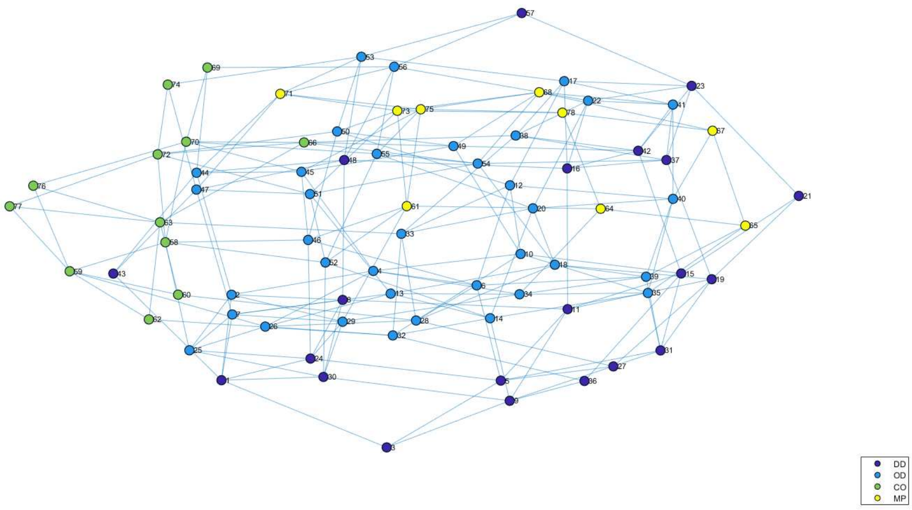

Figure A6: Graph of the similarity classes in G^* implied by the Level-1(α) rule. It does not separate any games. Each node is a game in G^* and a link is drawn between a pair of games if: (i) for some representation of the pair of games in G, both games are neighbors and (ii) both games belong to the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

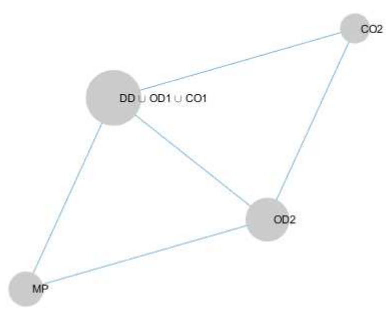

Figure A7: Graph of the Robinson-Goforth topology in G^*, with nodes representing the similarity classes according to Level-$k(\alpha)$ rules, $k = 1, \ldots, 5$, and edges how they connect.

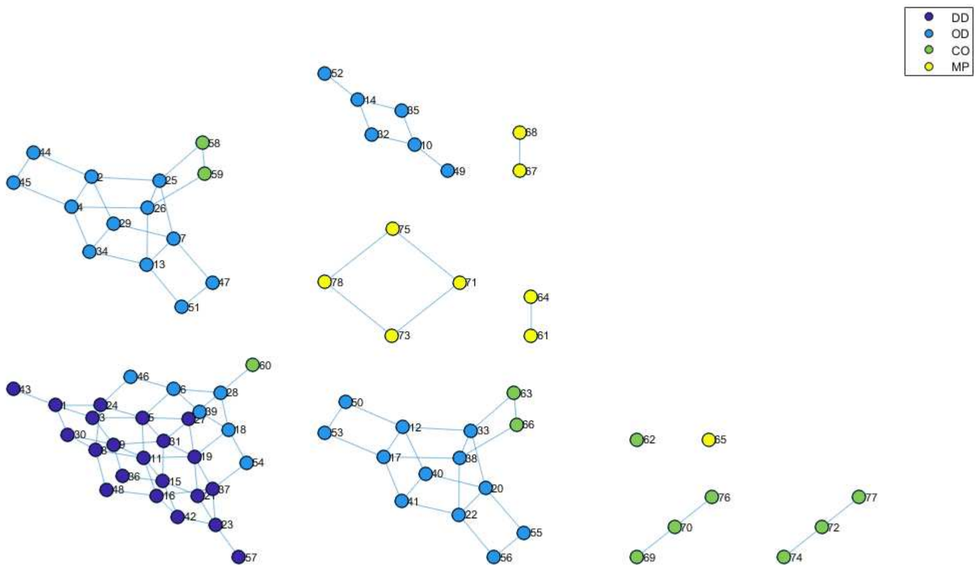

Figure A8: Graph of the similarity classes in G^* implied by the Level-k rule. Each node is a game in G^* and a link is drawn between a pair of games if: (i) for some representation of the pair of games in G, both games are neighbors and (ii) both games belong to the same similarity class.

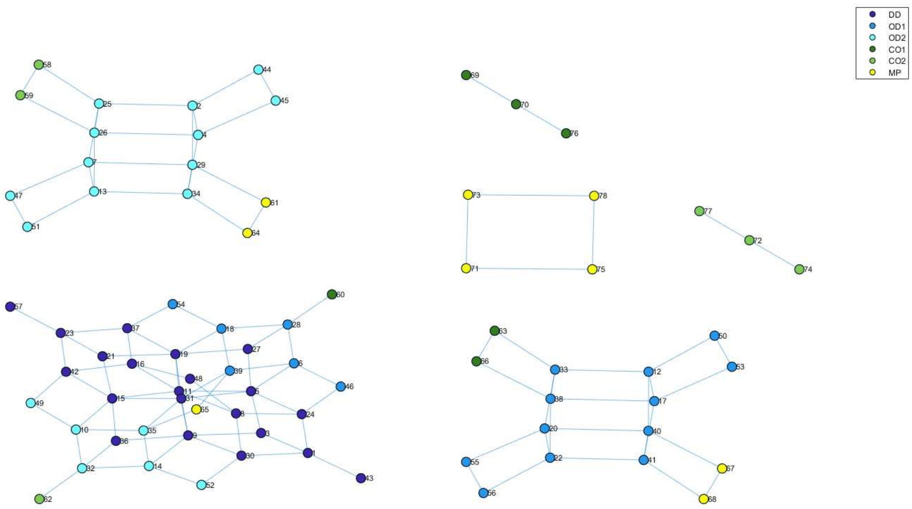

Figure A9: Graph of the similarity classes in G^* implied by the Level-1 rule. Each node is a game in G^* and a link is drawn between a pair of games if: (i) for some representation of the pair of games in G, both games are neighbors and (ii) both games belong to the same similarity class.

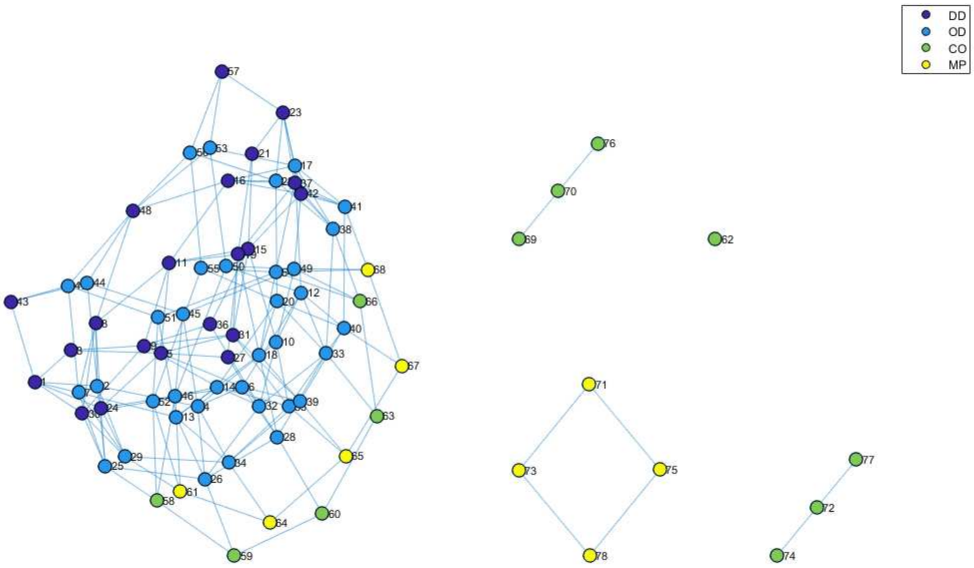

Figure A10: Graph of the similarity classes in G^* implied by the Level-2 rule. Each node is a game in G^* and a link is drawn between a pair of games if: (i) for some representation of the pair of games in G, both games are neighbors and (ii) both games belong to the same similarity class.

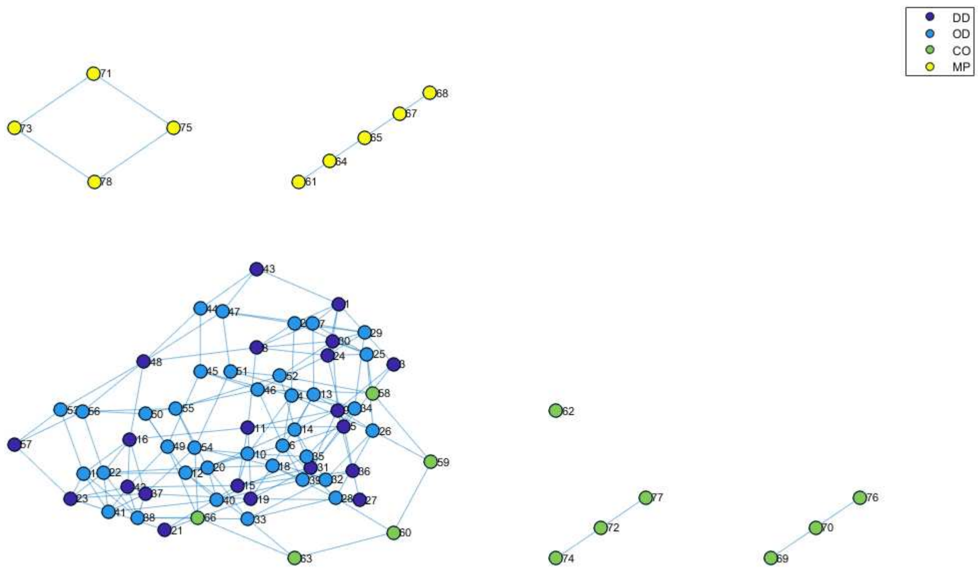

Figure A11: Graph of the similarity classes in G^* implied by the Level-4 rule. Each node is a game in G^* and a link is drawn between a pair of games if: (i) for some representation of the pair of games in G, both games are neighbors and (ii) both games belong to the same similarity class.

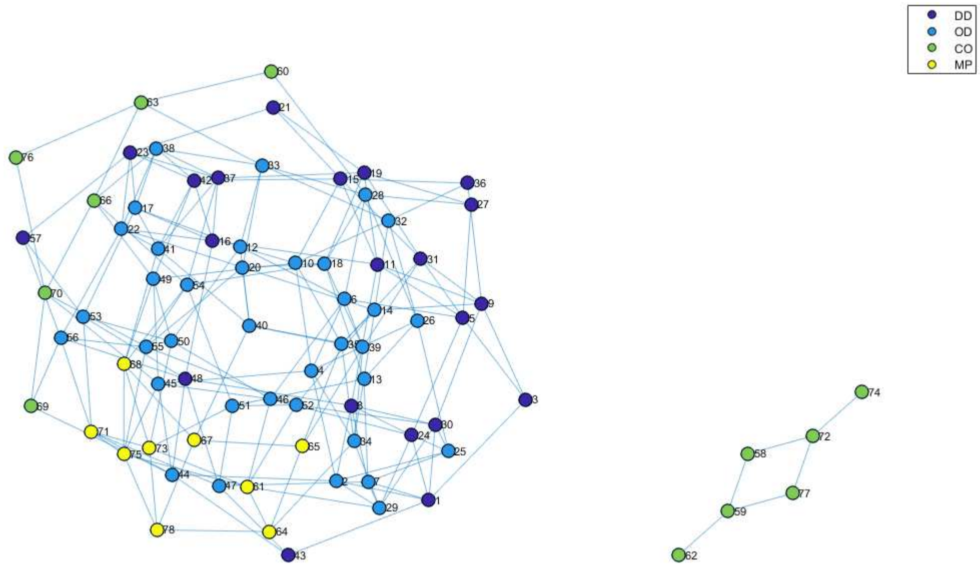

Figure A12: Graph of the similarity classes in G^* implied by the Level-5(α) rule. Each node is a game in G^* and a link is drawn between a pair of games if: (i) for some representation of the pair of games in G, both games are neighbors and (ii) both games belong to the same similarity class.

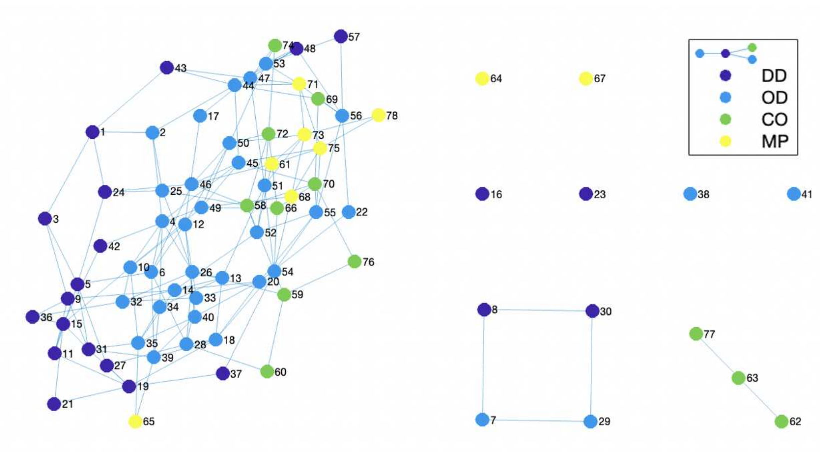

Figure A13: Graph of the similarity classes in G^* implied by self-favoring and other-favoring Near-Equal Split. Each node is a game in G^* and a link is drawn between a pair of games if: (i) for some representation of the pair of games in G, both games are neighbors and (ii) both games belong to the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

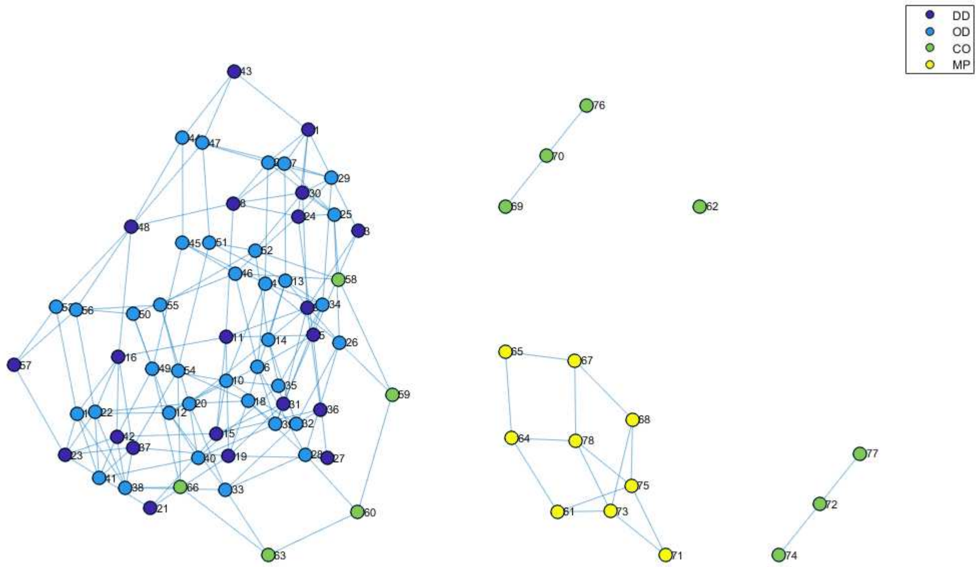

Figure A14: Graph of the similarity classes in G^* implied by Risk-Dominant Nash equilibrium. Each node is a game in G^* and a link is drawn between a pair of games if: (i) for some representation of the pair of games in G, both games are neighbors and (ii) both games belong to the same similarity class.

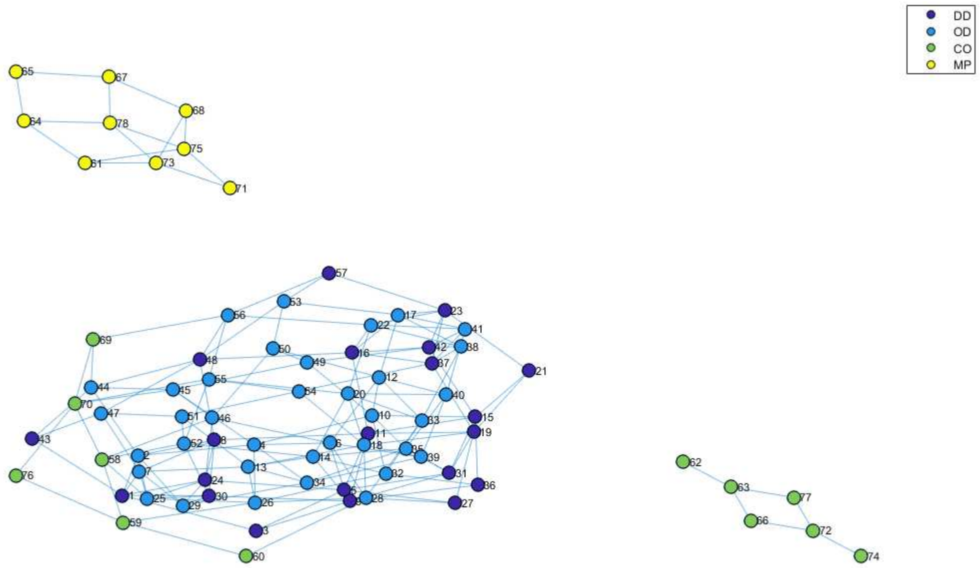

Figure A15: Graph of the similarity classes in G^* implied by Pareto-Dominant Nash equilibrium. Each node is a game in G^* and a link is drawn between a pair of games if: (i) for some representation of the pair of games in G, both games are neighbors and (ii) both games belong to the same similarity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

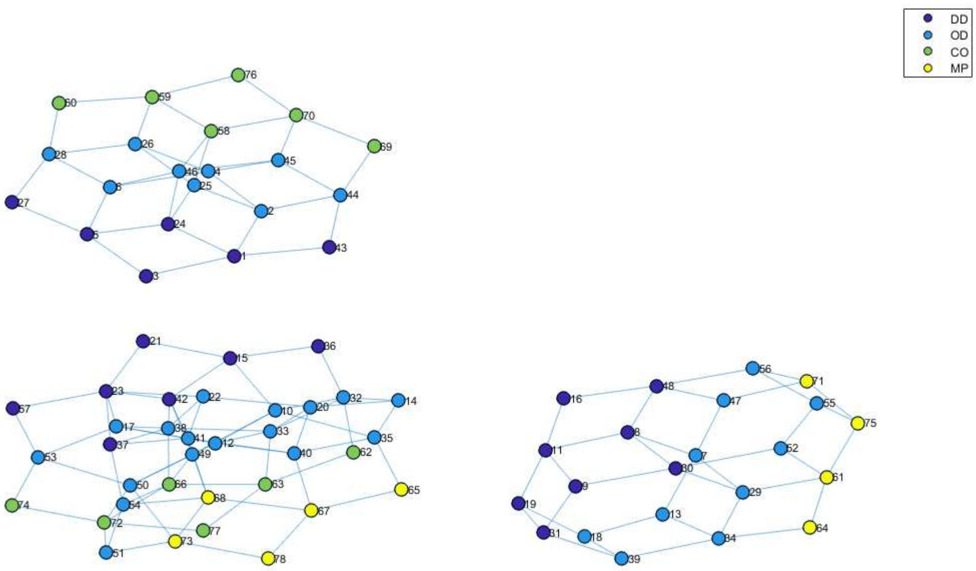

Figure A17: Graph of the similarity classes in G^* implied by Pareto Efficiency. Each node is a game in G^* and a link is drawn between a pair of games if: (i) for some representation of the pair of games in G, both games are neighbors and (ii) both games belong to the same similarity class.

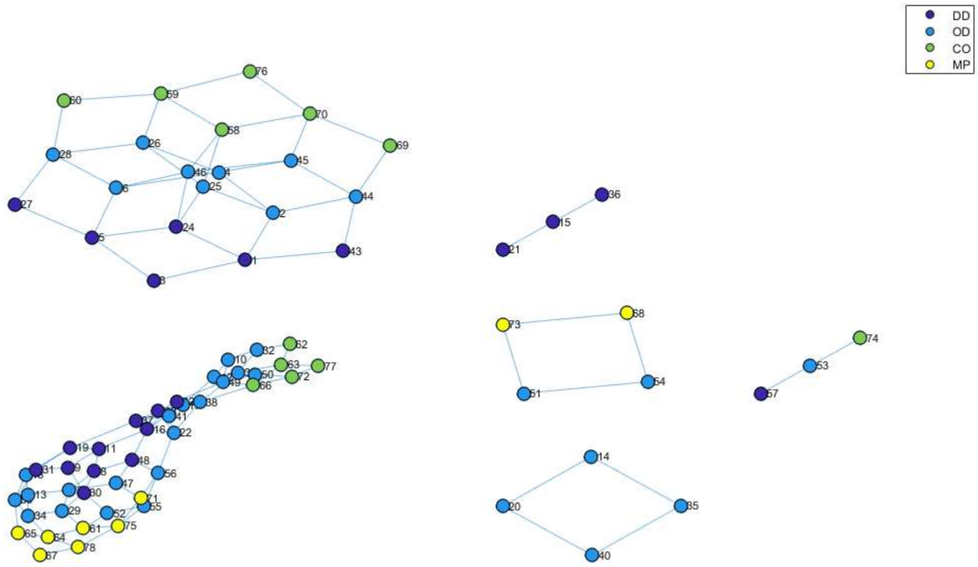

Figure A18: Graph of the similarity classes in G^* implied Equal Split. Each node is a game in G^* and a link is drawn between a pair of games if: (i) for some representation of the pair of games in G, both games are neighbors and (ii) both games belong to the same similarity class.

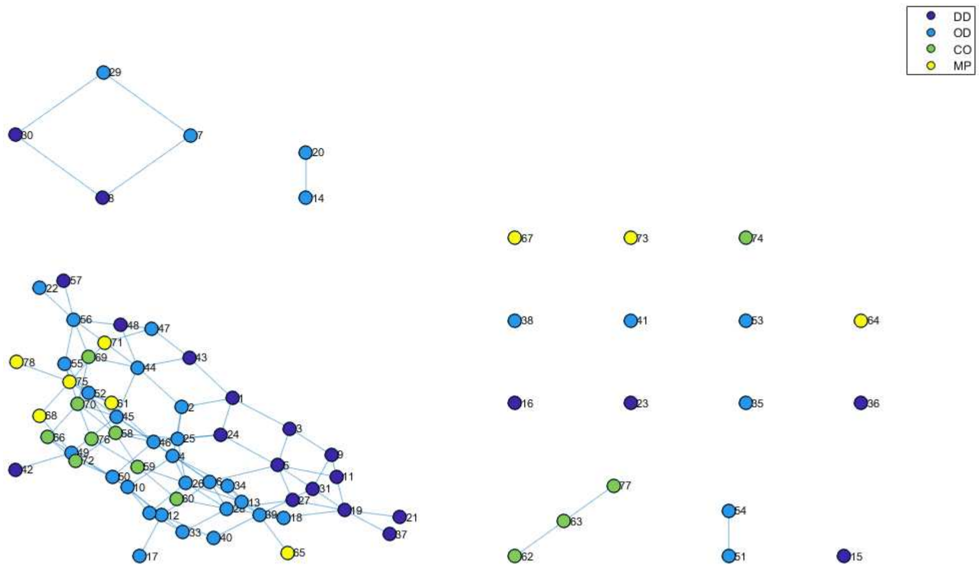

Figure A19: Graph of the similarity classes in G^* implied by the Soc-Max rule. Each node is a game in G^* and a link is drawn between a pair of games if: (i) for some representation of the pair of games in G, both games are neighbors and (ii) both games belong to the same similarity class.

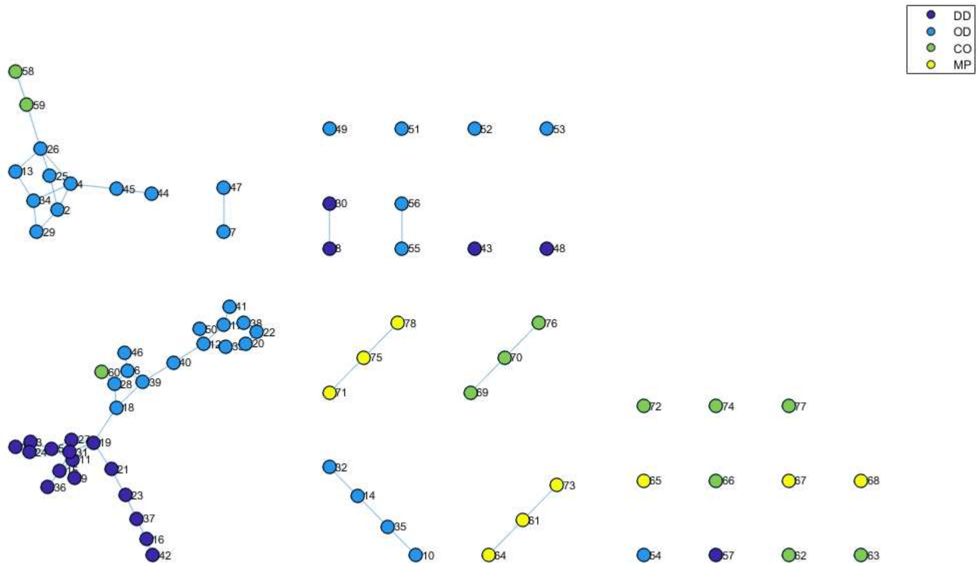

Figure A20: Graph of the empirical similarity classes in G^* from aggregate behavior of all subjects, for $\delta = 0\%$ at a 99.9% confidence level. Each node is a game in G^* and a link is drawn between a pair of games if: (i) for some representation of the pair of games in G, both games are neighbors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

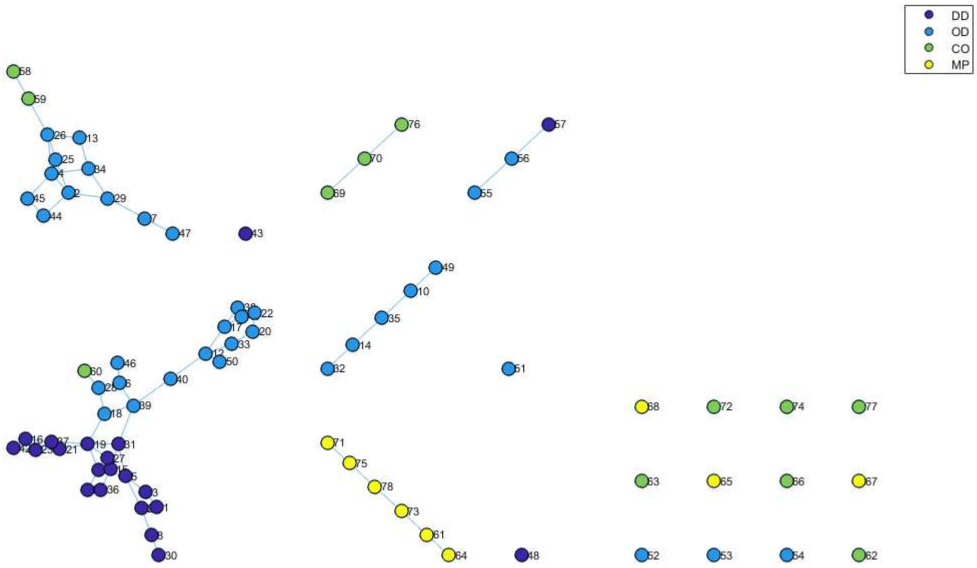

Figure A21: Graph of the empirical similarity classes in G^* from aggregate behavior of all subjects, for $\delta = 1\%$ at a 99.9% confidence level. Each node is a game in G^* and a link is drawn between a pair of games if: (i) for some representation of the pair of games in G, both games are neighbors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

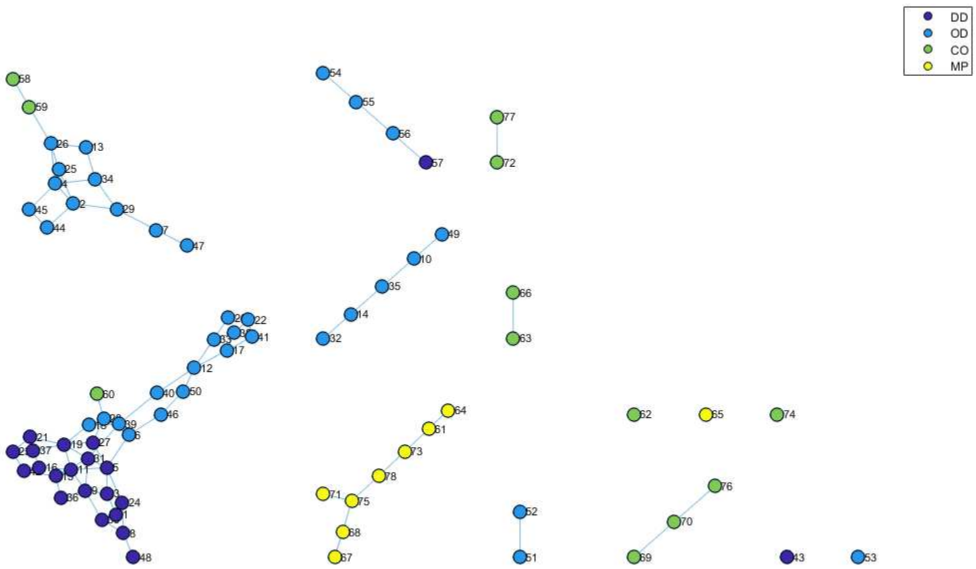

Figure A22: Graph of the empirical similarity classes in G^* from aggregate behavior of all subjects, for $\delta = 2\%$ at a 99.9% confidence level. Each node is a game in G^* and a link is drawn between a pair of games if: (i) for some representation of the pair of games in G, both games are neighbors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

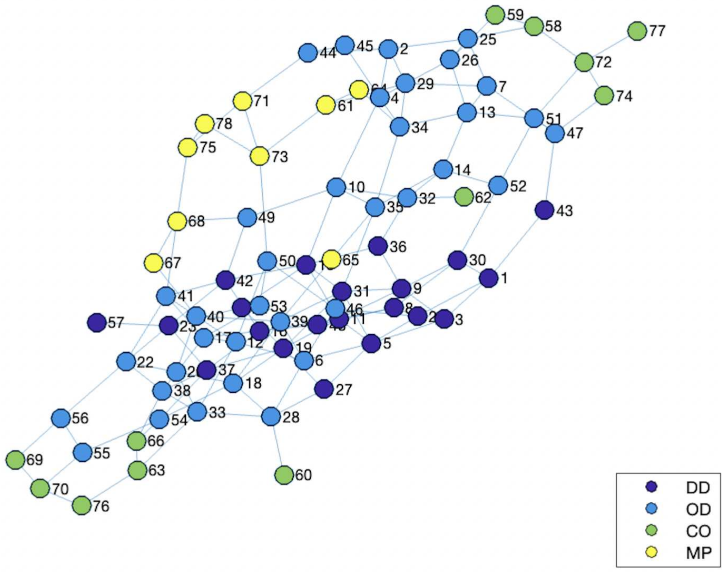

Figure A23: Graph of the empirical similarity classes in G^* from aggregate behavior of just those subjects that were identified as being level-1(α) types, for $(\Delta, p) = (5\%, 99.9\%)$.

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

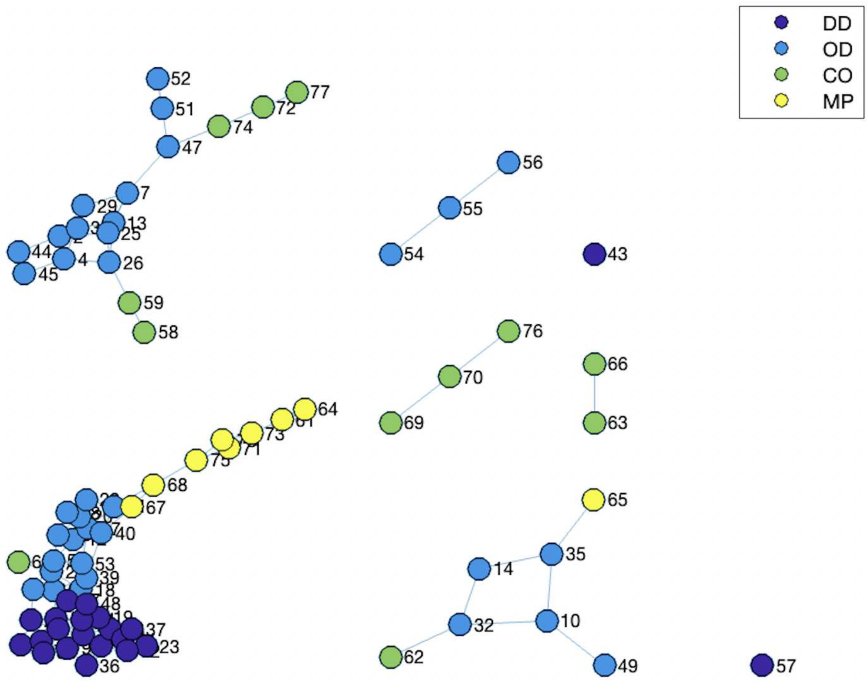

Figure A24: Graph of the empirical similarity classes in G^* from aggregate behavior of just those subjects that were identified as being level-$k(\alpha)$ types, for $k = 2, \ldots, 5$, and for $(\Delta, p) = (5\%, 99.9\%)$.

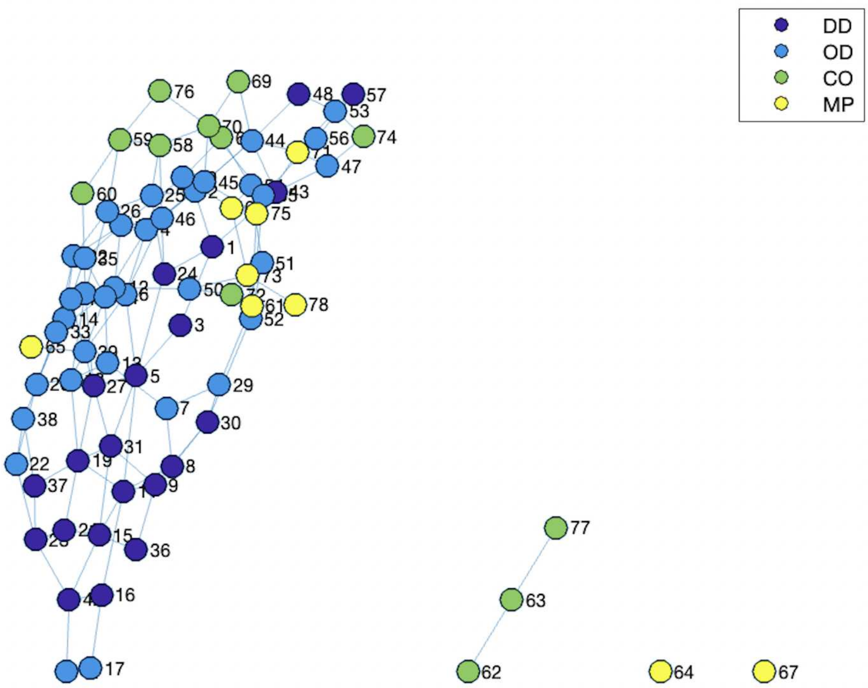

Figure A25: Graph of the empirical similarity classes in G^* from aggregate behavior of just those subjects that were identified as being cNES or rNES types, for $(\Delta, p) = (5\%, 99.9\%)$.

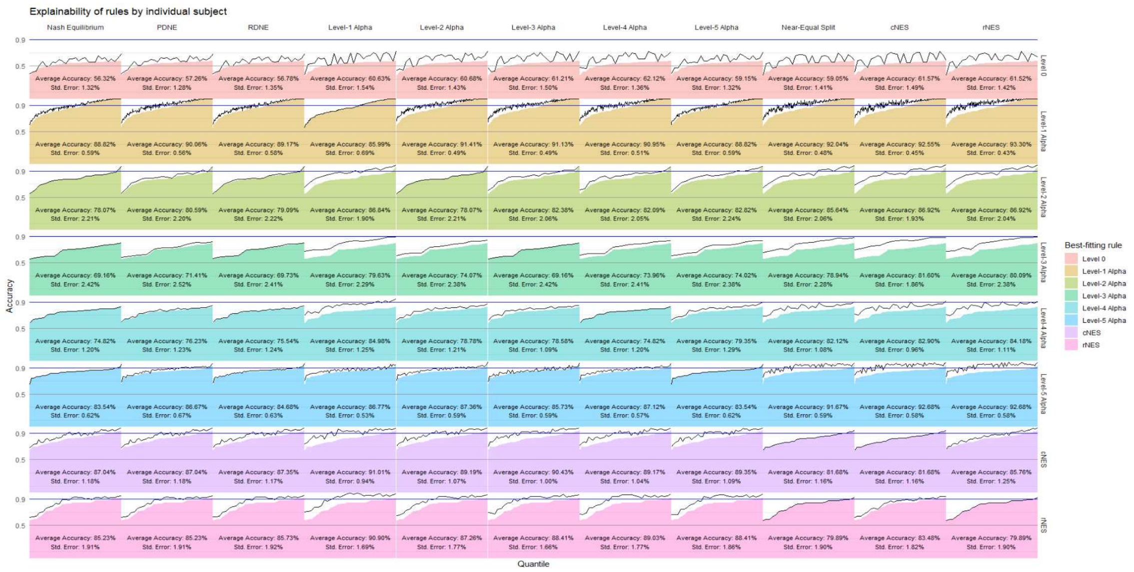

Figure A26: The quantile function of the accuracy of each subject's best-fitting rule (area in color) plus the added accuracy of other rules in games where the best-fitting rule predicts incorrectly. The rows correspond to the groups of subjects according to their best-fitting rule. The columns correspond 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: Transition matrix of rule switch: Numbers of subjects for each best-fitting primary rule (rows) by best-fitting secondary rule (columns),

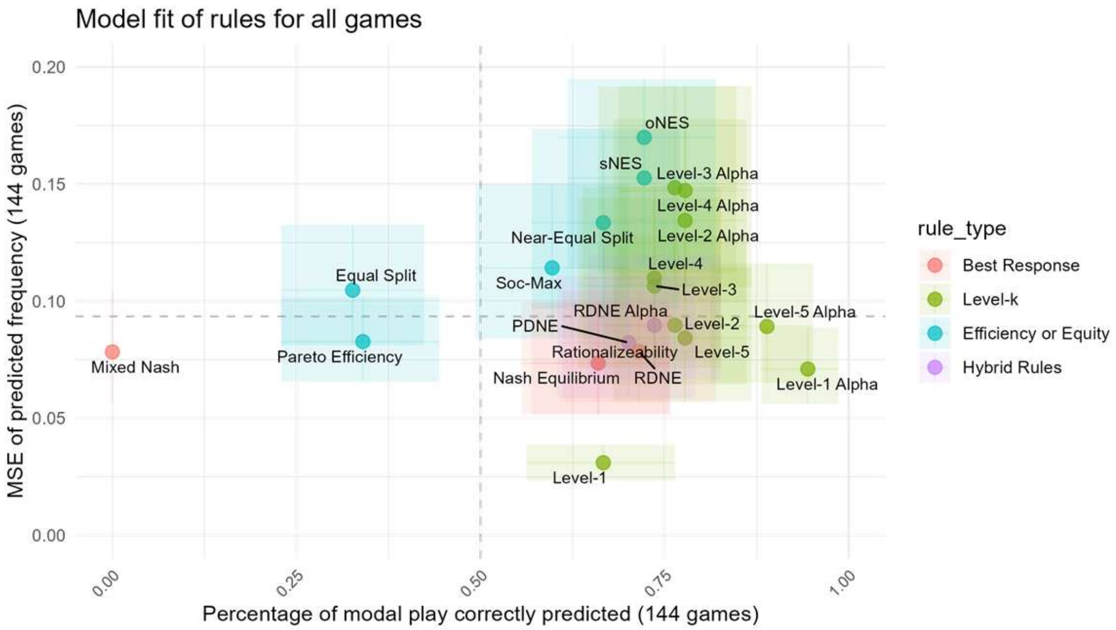

Figure A28: Model fit for behavior rules across all 144 perspectives. The x-axis corresponds to accuracy in modal play whereas the y-axis corresponds to MSE. The rectangles correspond to 99% confidence intervals of each statistic.

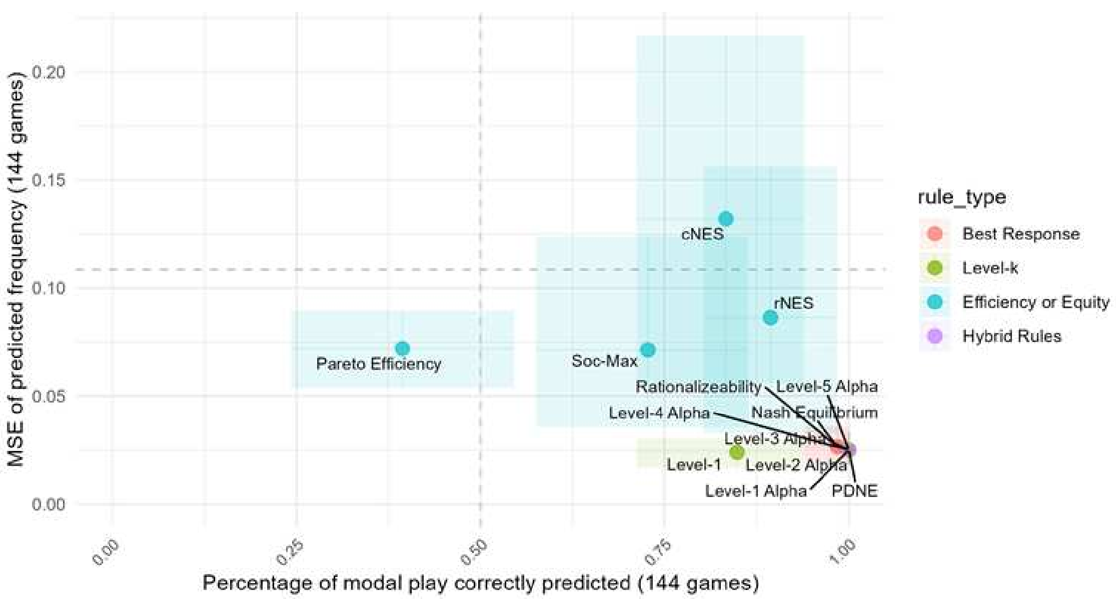

Figure A29: Model fit for behavior rules across the perspectives in the One-outcome empirical similarity class. The x-axis corresponds to accuracy in modal play whereas the y-axis corresponds to MSE. The rectangles correspond to 99% confidence intervals of each statistic.

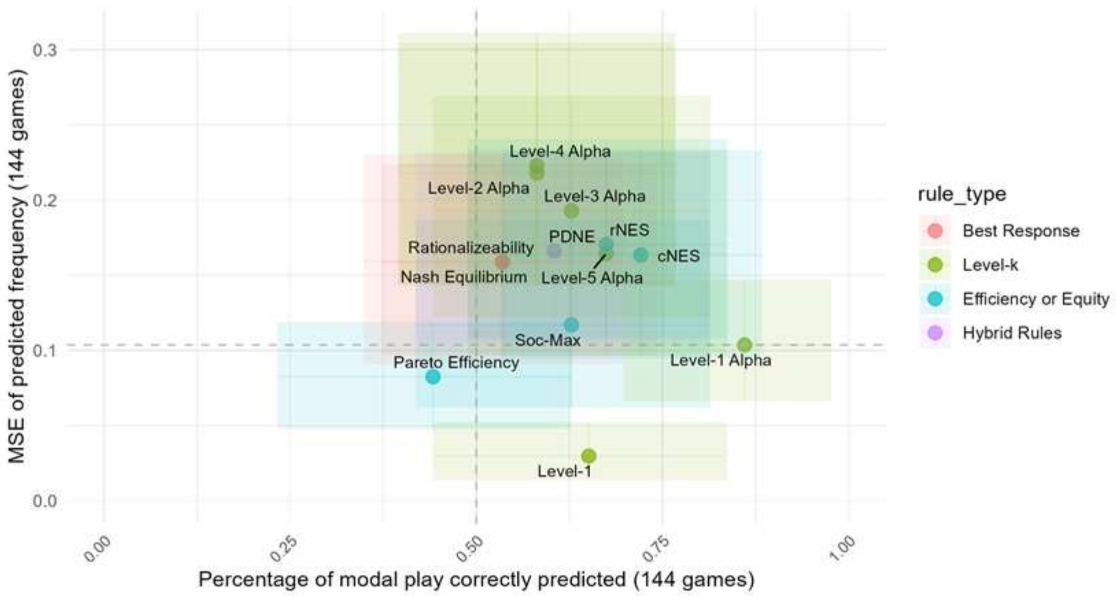

Figure A30: Model fit for behavior rules across the perspectives in the two outcome empirical similarity class. The x-axis corresponds to accuracy in modal play whereas the y-axis corresponds to MSE. The rectangles correspond to 99% confidence intervals of each statistic.

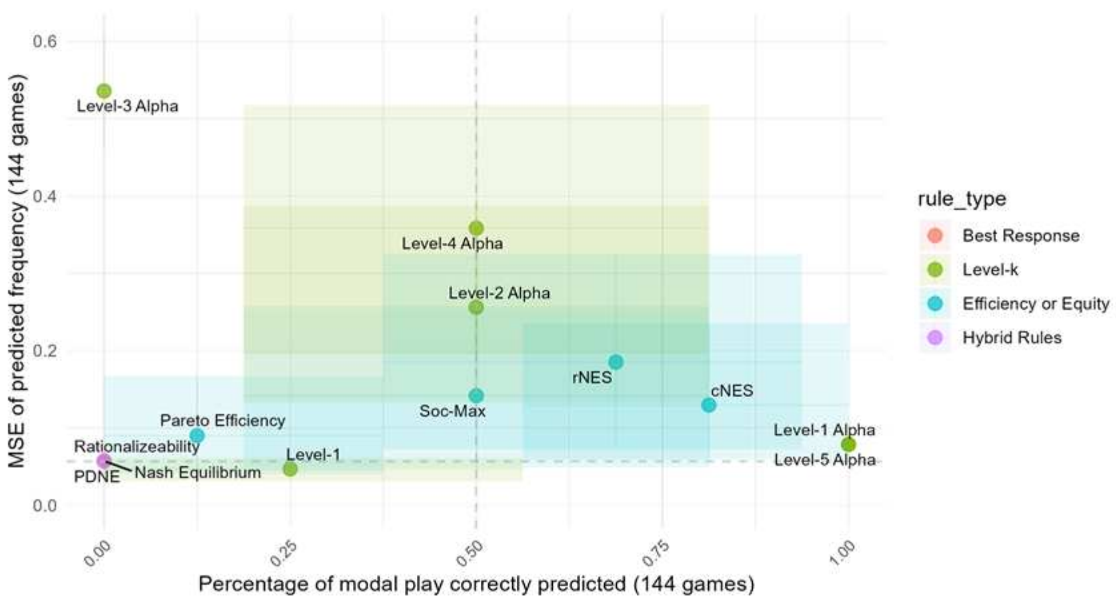

Figure A31: Model fit for behavior rules across the perspectives in the matching pennies empirical similarity class. The x-axis corresponds to accuracy in modal play whereas the y-axis corresponds to MSE. The rectangles correspond to 99% confidence intervals of each statistic.

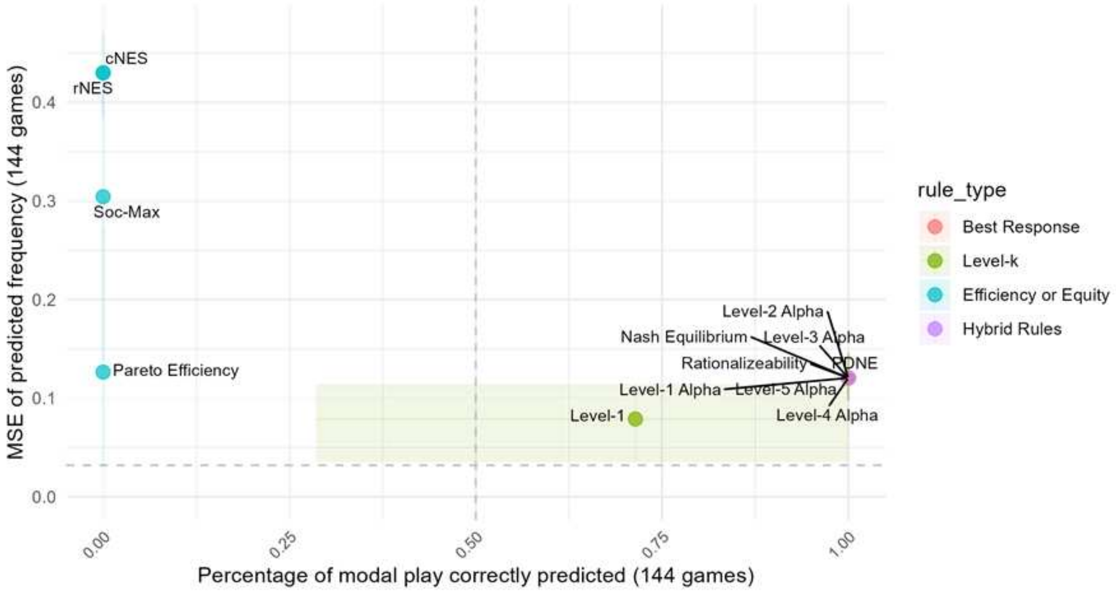

Figure A32: Model fit for behavior rules across the perspectives in the Prisoner's Dilemma empirical similarity class. The x-axis corresponds to accuracy in modal play whereas the y-axis corresponds to MSE. The rectangles correspond to 99% confidence intervals of each statistic.

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.[36]

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

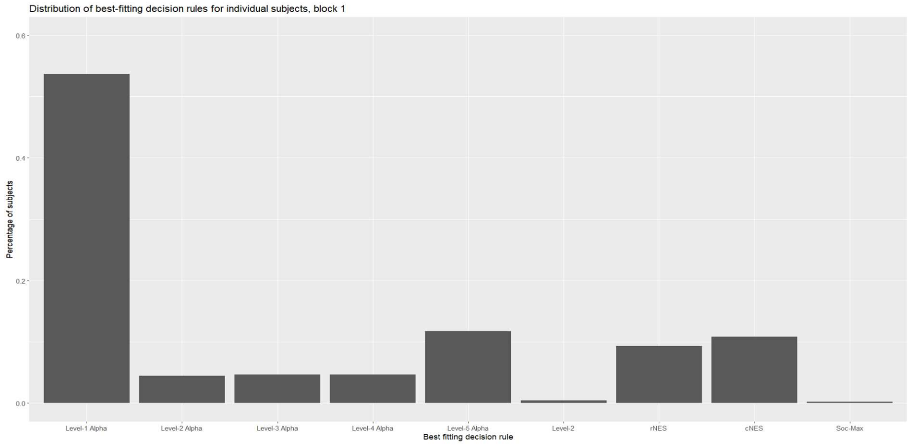

Figure A33: The distribution of best-fitting rules for our 450 subjects, using only the first 48 perspectives that subjects played.

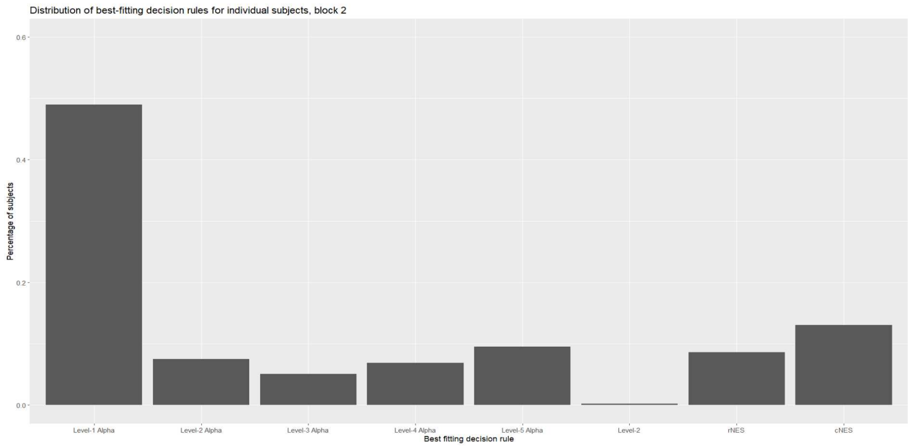

Figure A34: The distribution of best-fitting rules for our 450 subjects, using only the 49th to 95th perspectives that subjects played.

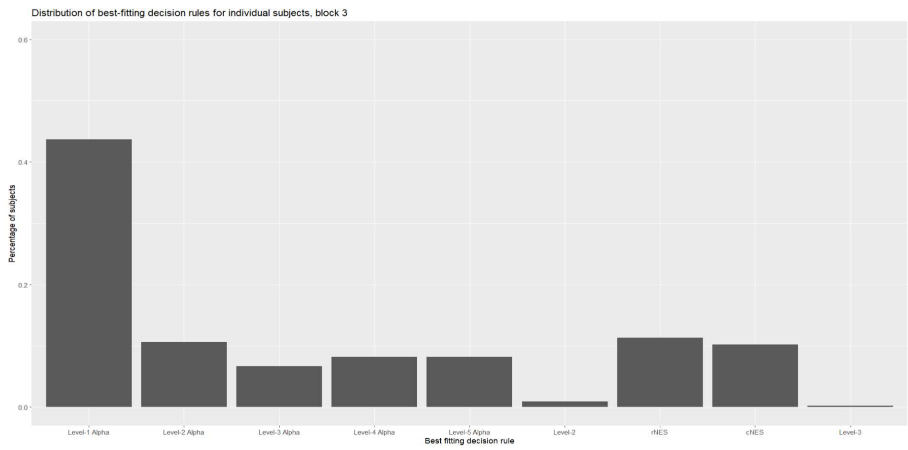

Figure A35: The distribution of best-fitting rules for our 450 subjects, using only the last 48 perspectives that subjects played.

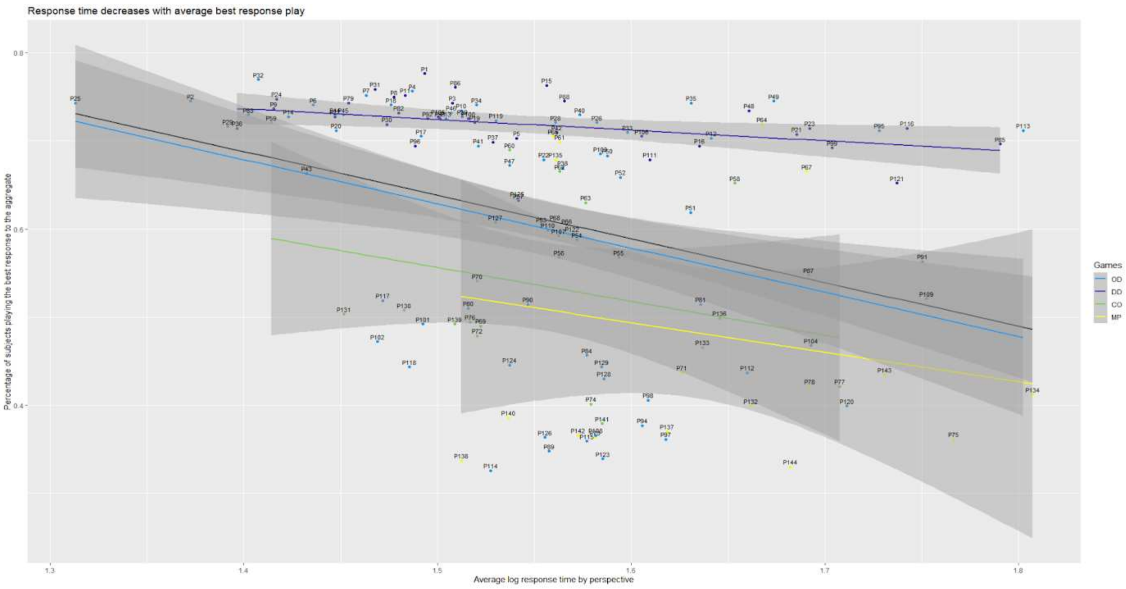

Figure A36: The relationship between log response time by perspective and the percentage of subjects playing the best response to the frequency distribution for each game.

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.

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.

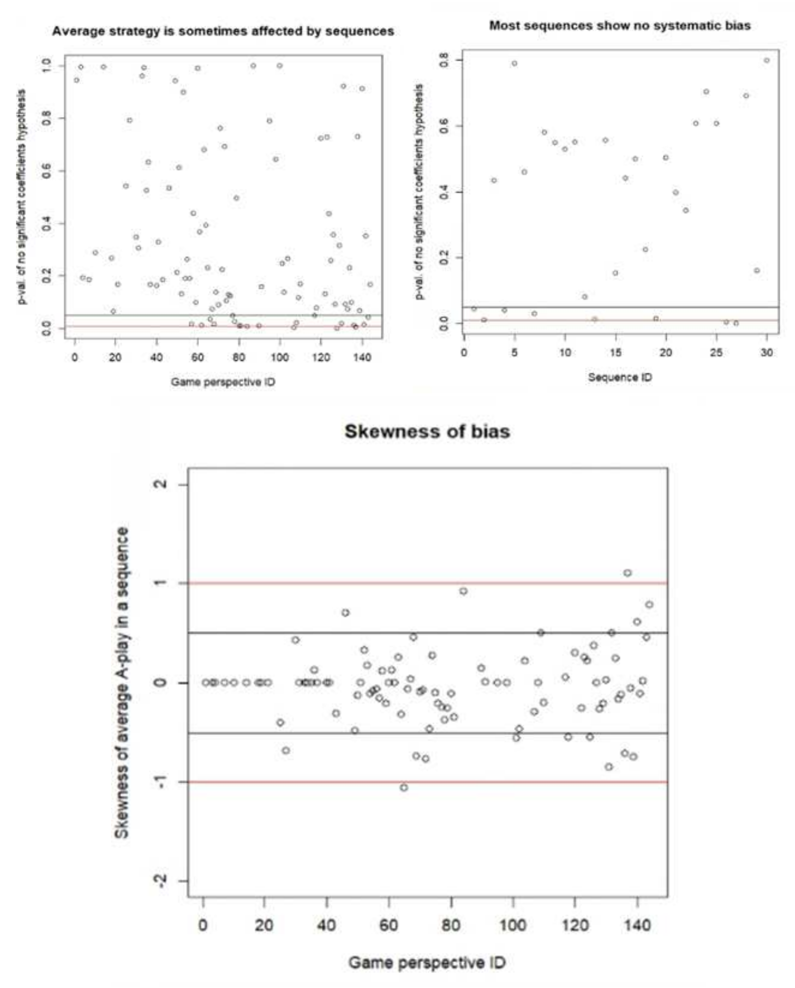

Figure A38: Summary of sequence effects in the all games experiment.

$\mathbf{g_{75}}$	A	B	
A	4, 1	1, 3	**0.24**
B	3, 4	2, 2	**0.76**
	0.39	**0.61**	

$\mathbf{g_{36}}$	A	B	
A	4, 2	$\underline{3}, \underline{3}$	**0.89**
B	1, 1	2, 4	**0.11**
	0.11	**0.89**	

$\mathbf{g_{61}}$	A	B	
A	4, 3	2, 4	**0.82**
B	1, 2	3, 1	**0.18**
	0.71	**0.29**	

Figure A39: g_{75}, g_{36}, and g_{61} with empirical relative frequencies in bold. Underlined payoffs indicate Nash equilibrium outcomes. The expected outcome of each game is (2.3,2.6), (3,3), (3,3.1).

D Mechanism design example

Consider a mechanism designer interested in altering g_{75} to induce an expected outcome of roughly $(3, 3)$. In our data, roughly 24% play $a_{1,1}$ and 39% play $a_{2,1}$ which gives an expected outcome of (2.3, 2.6). One of g_{75}'s neighbors achieves this goal (g_{61}) with 82% playing $a_{1,1}$ and 71% playing $a_{2,1}$, leading to an expected outcome of (3.0, 3.1). It achieves this by shifting the level-1(α) prediction from (2,2) in g_{75} to (3,4) in g_{61}. One of the standard mechanism design solutions, on the other hand, would be to design a game in which the desirable outcome $(3, 3)$ is the unique equilibrium. It can be verified that only $g_{14}, g_{15}, g_{20}, g_{21}, g_{35}, g_{36}$, and g_{40} satisfy this condition. However, in our data, this is only achieved by the best case scenario in which $(3, 3)$ is implemented by strictly dominant action (in g_{36} where 89% play the equilibrium strategy). In all other games with a unique (3,3) equilibrium, the percentage of equilibrium play is strictly lower, with consequently worse (or not Pareto improving) expected outcomes. Moreover, none of the games with a unique (3,3) equilibrium are neighbors with g_{75}. In other words, altering g_{75} into g_{36} requires larger payoff changes (because g_{36} is not a neighbor to g_{75}) and achieves a worse expected outcome ((3,3)) than our suggested solution of altering g_{75} into g_{61} ((3,3.1)).

E 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 where Level-$k(\alpha)$ rules span all outcomes and hence both NE (CO2).

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 1 4 2 3	4,6 7,9	OD1	A1
42	4 1 3 2 1 3 2 4	4,7 6,9	DD	A1
43	4 4 2 3 3 2 1 1	12,1	DD	A1
44	4 4 2 3 3 1 1 2	7,1 12,6	OD2	A2
45	4 4 2 2 3 1 1 3	8,1 12,5	OD2	A2
46	4 4 2 1 3 2 1 3	9,1 12,4	OD1	A1
47	4 3 2 4 3 2 1 1	1,1 12,12	OD2	A2
48	4 3 2 4 3 1 1 2	6,1 12,7	DD	A1
49	4 3 2 2 3 1 1 4	3,7 6,10	OD2	A2
50	4 3 2 1 3 2 1 4	2,7 6,11	OD1	A1
51	4 2 2 4 3 3 1 1	2,1 12,11	OD2	A2
52	4 2 2 3 3 4 1 1	6,4 9,7	OD2	A2
53	4 2 2 1 3 3 1 4	1,7 6,12	OD1	A1
54	4 1 2 4 3 3 1 2	3,1 12,10	OD1	B1
55	4 1 2 3 3 4 1 2	6,5 8,7	OD1	B1
56	4 1 2 2 3 4 1 3	6,6 7,7	OD1	B1
57	4 1 2 2 3 3 1 4	6,7	DD	B1
58	4 4 2 3 1 1 3 2	7,4 9,6	CO2	A2
59	4 4 2 2 1 1 3 3	8,4 9,5	CO2	A2
60	4 4 2 1 1 2 3 3	9,4	CO1	A1

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		