

Adaptive Noise in Human Game Play: An Information-Theoretic Signature that Separates Learning Models on Matching-Pennies Data

Pratyush Mahadevaiah

`pratyush.mahadevaiah@gmail.com`

Dougherty Valley High School

Research Article

Keywords: learning in games, transfer entropy, experience-weighted attraction, reinforcement learning, matching pennies, bounded rationality

Posted Date: October 21st, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10031328/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Adaptive Noise in Human Game Play: An Information-Theoretic Signature that Separates Learning Models on Matching-Pennies Data

[Author name removed for double-blind review] [Affiliation removed for double-blind review]

Abstract

Reinforcement learning is a standard model of how humans play repeated games, yet it predicts that exploration anneals away and play becomes deterministic. Human subjects do not behave this way. We characterize human play in the Selten and Chmura (2008) 2×2 matching-pennies experiments with four information-theoretic signatures of the action trajectory, computed on period-by-period data from all 108 independent subject groups. Two signatures dominate. Humans maintain a high, roughly stationary policy entropy across the whole 200-period session (mean entropy rate 0.64 bits), and they show strong transfer entropy from the opponent’s last action to their own next action (0.06 bits; 55% of the 864 individual subjects exceed a 100-shuffle surrogate null at the 5% level). Matched Q-learning reproduces neither: it anneals to near-deterministic play (entropy rate 0.06) and carries essentially no opponent-responsiveness (transfer entropy 0.002; 7% of agents exceed the null, barely above the false-positive rate). No temperature setting closes both gaps at once, because Q-learning trades stochasticity against responsiveness while humans achieve both. Running four further standard learning models through the identical protocol, only self-tuning EWA enters the human region of the (entropy rate, transfer entropy) plane. This recovers, from a model-free trajectory measure, the model-selection result of Chmura, Goerg and Selten (2012), who found self-tuning EWA best predicts individual round-by-round choices. The transfer-entropy signature is consistent with this: among standard learners, only self-tuning EWA is simultaneously stochastic and opponent-responsive, as humans are. We are explicit that this is an architectural correlation rather than a demonstrated cause.

Keywords: learning in games, transfer entropy, experience-weighted attraction, reinforcement learning, matching pennies, bounded rationality

JEL Classification: C72, C73, C92, D83

1. Introduction

People learning to play a mixed-strategy game face a tension. They must keep mixing, because any predictable pattern is exploited, yet they must also respond to what opponents do. Standard reinforcement learning resolves this tension in the wrong direction: as value estimates sharpen and exploration decays, the policy concentrates and play becomes deterministic. On the Selten and Chmura (2008) matching-pennies games this is visibly false. Subjects keep mixing for 200 periods and never settle on a pure action.

We study this directly with information theory. Rather than fit a learning model and read off parameters, we compute four signatures of the raw action trajectory and ask which features of human play any model must reproduce. The signatures follow the “adaptive noise” framework of an earlier companion paper by the present author Author (2025): the oscillation amplitude of the policy entropy, fluctuations of the policy’s divergence from uniform, the decline of policy entropy over the session, and the transfer entropy between players Schreiber (2000). The first three describe how a player’s own randomness evolves; the fourth describes whether that randomness carries information about the opponent.

The dataset has a substantial model-comparison literature, almost all of it concentrated on stationary, aggregate predictions. Selten and Chmura (2008) compared Nash equilibrium against quantal response equilibrium McKelvey and Palfrey (1995), action-sampling, payoff-sampling, and impulse balance equilibrium on long-run action frequencies; Brunner et al. (2011) re-examined that comparison, and the original authors replied with an analysis of round-by-round behavior Selten et al. (2011). Chmura et al. (2012) then compared learning models on individual choice prediction. Our approach is complementary: instead of asking which model predicts frequencies or next actions best, we ask which dynamic properties of the trajectory a model must possess at all. Transfer entropy has been used to detect directed information flow in financial time series Marschinski and Kantz (2002); to our knowledge it has not been used to discriminate learning models on experimental game data.

Our contribution is threefold. First, on period-by-period data from all 108 subject groups, we show that human play has two robust signatures that matched Q-learning fails to reproduce: maintained high policy entropy and strong opponent-to-self transfer entropy. Second, we show the failure is structural, not a matter of tuning: across the full temperature range, Q-learning trades raw stochasticity against opponent-responsiveness and never reaches the region humans occupy. Third, running four further standard learning models Erev and Roth (1998); Fudenberg and Levine (1998); Ho et al. (2007); Chmura et al. (2012) through the same protocol, only self-tuning EWA enters the human region. This converges with Chmura

et al. (2012), who found self-tuning EWA best describes individual round-by-round behavior in this dataset by a choice-prediction criterion. Our trajectory measure reaches the same model by an unrelated route and identifies the structural property the winning model shares with human play: only self-tuning EWA is at once noisy and adaptive. We treat this as a diagnostic correlation between model architecture and a measurable trajectory signature, not as a causal account of why EWA predicts choices well. More broadly, the contribution is a diagnostic dimension rather than a new model: instead of asking which model predicts choices best, we ask which dynamical signatures any successful model must reproduce, and screen models on that basis before any fitting.

2. Experimental data

Selten and Chmura (2008) ran twelve generalized matching-pennies games. Games 1–6 are constant-sum; games 7–12 are non-constant-sum transformations of games 1–6 that preserve the best-response structure. Every game is a 2×2 game with a unique, completely mixed Nash equilibrium and cyclic best responses, so no pure-strategy resting point exists and sustained mixing is the equilibrium prediction. The mixed equilibria span a wide range: the Nash probability of the row player’s first action runs from 0.09 (games 1 and 7) to 0.45 (games 6 and 12), which lets us ask how human behavior varies as the equilibrium moves from the edge of the strategy simplex toward its center.

Each game was played by independent subject groups of eight: four row players and four column players in fixed roles, randomly rematched within the group in every one of 200 periods. The design has 12 independent groups per constant-sum game and 6 per non-constant-sum game, for 108 groups and 864 subjects in total. Full laboratory procedures, instructions, and payment details are given in Selten and Chmura (2008).

We obtained the period-by-period records from the original authors: 172,800 subject-period observations, each recording the subject’s action, the action of the opponent faced that period, and the opponent’s identity under the random matching. We verified the data three ways: the action coding was confirmed against the role-specific frequency structure; the per-period pairings were confirmed to form valid matchings with mutually consistent opponent records; and every subject’s recovered 200-period action frequency was confirmed to reproduce the corresponding aggregate in the publicly archived release of the dataset. The public archive contains only aggregate frequencies, which is why the trajectory analysis below requires the period-level records.

3. Method

3.1 Signatures

For a player we estimate the time-varying mixing probability by a sliding window of $w = 20$ periods over their own actions, giving a policy-entropy trajectory H_t in bits. On H_t we compute three signatures. The *oscillation amplitude* is the peak of the discrete Fourier spectrum of the linearly detrended H_t . The *KL fluctuation* is the detrended standard deviation of $1 - H_t$, the divergence of a binary policy from uniform. The *entropy decline* is the linear slope of H_t over the session; a strongly negative slope is annealing toward determinism, a slope near zero is maintained noise. Estimator details are in Appendix A; results are insensitive to w in the range 15–30.

The fourth signature is the *transfer entropy* from the opponent’s action to the player’s own next action Schreiber (2000),

$$T_{\text{opp} \rightarrow \text{own}} = H(a_t \mid a_{t-1}) - H(a_t \mid a_{t-1}, o_{t-1}), \quad (1)$$

where a_t is the player’s action and o_{t-1} the action of the opponent they faced in the previous period. This measures how much the opponent’s last move reduces uncertainty about the player’s next move, beyond the player’s own history. We estimate the entropies with the Grassberger finite-sample correction Grassberger (2008) and subtract a 50-shuffle surrogate baseline Marschinski and Kantz (2002) to remove residual bias. All entropies use base-2 logarithms.

Because the headline contrast is dramatic and the estimator might be questioned, we also report a separate *entropy rate*, the order-1 Markov conditional entropy of each subject’s action sequence, $H(a_t \mid a_{t-1})$. This is the single-number summary of how stochastic a subject’s play is, and it is computed without reference to the opponent.

3.2 Learning models

We compare human play against five learning models. Q-learning uses Boltzmann exploration with annealing temperature. Reinforcement learning Erev and Roth (1998) reinforces the chosen action by its realized payoff. Stochastic fictitious play Fudenberg and Levine (1998) logit-best-responds to the running empirical frequency of the opponent. Self-tuning EWA Ho et al. (2007) updates attractions to all actions using foregone payoffs, with a self-tuning change-detector that reweights recent versus cumulative experience. Impulse-matching learning Chmura et al. (2012), in the spirit of Selten’s learning direction theory Selten and Buchta (1999), computes impulses on a transformed payoff matrix in which

shortfalls below the pure-strategy maximin are counted double, and chooses actions with probability proportional to cumulated impulse. Full update equations and parameter values for every model are in Appendix B.

3.3 Simulation design

Each model is simulated under the experimental protocol exactly: four row and four column agents per group, a fresh uniformly random matching every period, 200 periods, on each game’s actual payoff matrix, with the number of simulated groups matched to the human design (12 per constant-sum game, 6 per non-constant-sum game, 864 agents per model). Each agent updates from the action of the opponent it actually faced, so the transfer-entropy construction is identical for humans and models. All models use their standard, untuned parameterizations (Table 1); parameter sensitivity is examined in Sections 4.4 and 4.7.

Table 1. Simulation protocol and model parameters.

Protocol	4 row + 4 column agents, 200 periods, random rematching groups matched to data: 12×6 CS, 6×6 NCS (864 agents/model)
Q-learning	$\alpha = 0.1$, $\gamma = 0.95$, $\tau_0 = 1$, decay $\rho = 0.99$
Reinforcement	initial propensity 1
Fictitious play	logit precision $\beta = 1$, uniform prior weight 1
Self-tuning EWA	$\lambda = 1$, $N(0) = 1$, $\kappa = 0$
Impulse-matching	initial impulse 1, pure cumulation ($\phi = 0$)

3.4 Statistical methodology

The primary unit of analysis is the independent subject group ($n = 108$); group-level signatures average over the eight members. Group-level comparisons use one-sample t -tests against zero, two-sample Kolmogorov–Smirnov tests, and Cohen’s d . Individual-level significance of transfer entropy is assessed against a 100-permutation surrogate null per subject at the 5% level. Because subjects are nested in groups and games, we additionally fit linear mixed-effects models on per-subject signatures with a random intercept for game, contrasting humans against each model (864 subjects per condition); coefficient estimates with confidence intervals are reported in Appendix C. Robustness of the transfer-entropy estimator to the entropy estimator and to the order of own-history conditioning is reported in Section 4.6.

4. Results

4.1 Humans stay stochastic; Q-learning anneals

Human individual policy entropy stays near 0.6 bits across the whole session, with only a slight decline (Figure 1). Q-learning collapses to near-zero entropy within about thirty periods. The entropy rate makes the gap explicit: 0.639 bits for humans against 0.057 for matched Q-learning (Cohen’s $d = 5.0$), and the distributions do not overlap at all — the least stochastic human group (0.249 bits) sits well above the most stochastic Q-learning group (0.136 bits). Maintaining noise is the first thing a model of this data must do, and decaying-exploration Q-learning does the opposite.

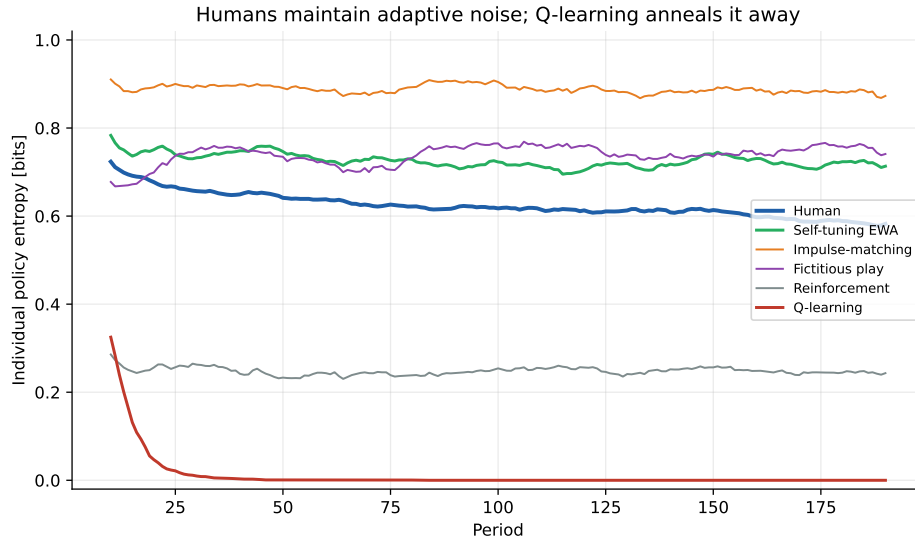


Figure 1. Mean individual policy entropy over the session. Humans hold near 0.6 bits; Q-learning anneals to zero within thirty periods; self-tuning EWA tracks the human level. Reinforcement concentrates at low entropy, while fictitious play and impulse-matching stay near maximal randomness.

4.2 Humans respond to opponents; Q-learning does not

Human transfer entropy from opponent to own next action averages 0.059 bits across the 108 groups ($t(107) = 15.8$, $p = 8 \times 10^{-30}$ against zero). Matched Q-learning averages 0.002 (Cohen’s $d = 2.07$); 86% of human groups exceed even the single most responsive Q-learning group. At the individual level, 55.0% of the 864 human subjects show transfer entropy above the 95th percentile of their own 100-shuffle surrogate null, against 7.3% of Q-learning agents — barely above the 5% false-positive rate (Table 5). Figure 2 shows the full group-level distributions: humans separate cleanly from Q-learning on every signature, with near-complete separation on transfer entropy and entropy rate.

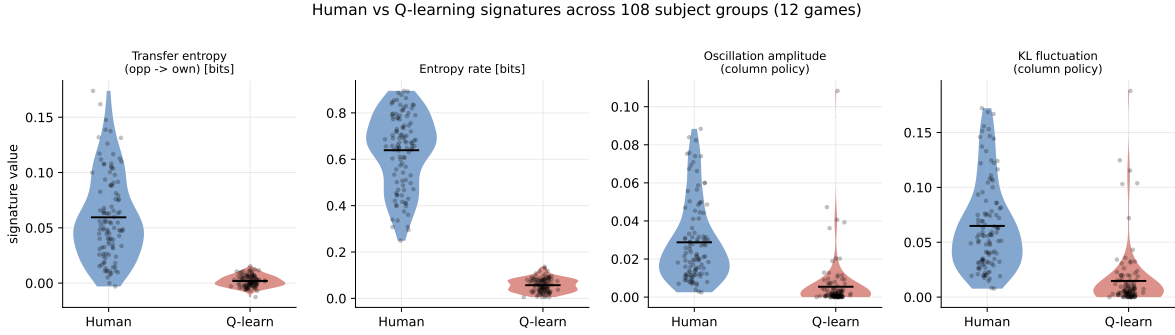


Figure 2. Human versus matched Q-learning signatures across all 108 groups. Humans exceed Q-learning on every signature, with near-complete separation on transfer entropy and entropy rate.

4.3 Responsiveness tracks the location of the equilibrium

Human transfer entropy is not uniform across games: it rises systematically as the mixed equilibrium moves toward the center of the strategy simplex (Table 2, Figure 3). Within the constant-sum block, mean transfer entropy rises from 0.023 bits in game 1 (Nash $p_U = 0.09$) to 0.103 bits in game 6 (Nash $p_U = 0.45$); the non-constant-sum block shows the same gradient (0.017 to 0.093). Across the twelve games, mean human transfer entropy correlates with equilibrium centrality, $-|p_U - 1/2|$, at Spearman $\rho = 0.81$ ($p = 0.001$). Entropy rate shows the same gradient, as it must when the equilibrium mixture itself moves toward uniform. The constant-sum versus non-constant-sum distinction, by contrast, has no detectable effect on either signature. Q-learning’s transfer entropy stays in the noise floor in every game, so the human cross-game pattern has no counterpart to match against (its rank correlation with the human pattern is $\rho = 0.29$, $p = 0.37$); we report this as a negative result whose cause is the absence of any responsiveness signal in the model, not the absence of structure in the human data.

4.4 The gap is structural, not a tuning artifact

The entropy gap could in principle be an artifact of the temperature schedule. It is not. Sweeping Q-learning temperature, including constant (non-annealing) temperature, entropy rate rises monotonically and passes through the human value, but transfer entropy peaks near 0.009 bits and then falls as the policy approaches uniform (Figure 4). At the temperature that matches human entropy rate, Q-learning transfer entropy is 0.004 bits, about 7% of the human value. Q-learning trades stochasticity against responsiveness: cold play is deterministic, hot play is uniform noise that ignores the opponent. Humans break the trade-off, sitting in a region of the plane Q-learning never reaches.

Table 2. Per-game human signatures (group mean $\pm$ SD) and the Nash mixed equilibrium. Games 1–6 constant-sum ($n = 12$ groups each), 7–12 non-constant-sum ($n = 6$ each). Transfer entropy rises as the equilibrium approaches the center of the simplex.

Game	Nash p_U	n	Transfer entropy (bits)	Entropy rate (bits)
1	0.09	12	0.023 ± 0.016	0.403 ± 0.080
2	0.18	12	0.052 ± 0.023	0.661 ± 0.071
3	0.27	12	0.027 ± 0.016	0.484 ± 0.089
4	0.36	12	0.059 ± 0.015	0.710 ± 0.061
5	0.36	12	0.084 ± 0.032	0.728 ± 0.142
6	0.45	12	0.103 ± 0.046	0.785 ± 0.072
7	0.09	6	0.017 ± 0.013	0.487 ± 0.150
8	0.18	6	0.082 ± 0.043	0.705 ± 0.078
9	0.27	6	0.038 ± 0.015	0.500 ± 0.073
10	0.36	6	0.067 ± 0.041	0.705 ± 0.068
11	0.36	6	0.077 ± 0.031	0.771 ± 0.056
12	0.45	6	0.093 ± 0.034	0.789 ± 0.140

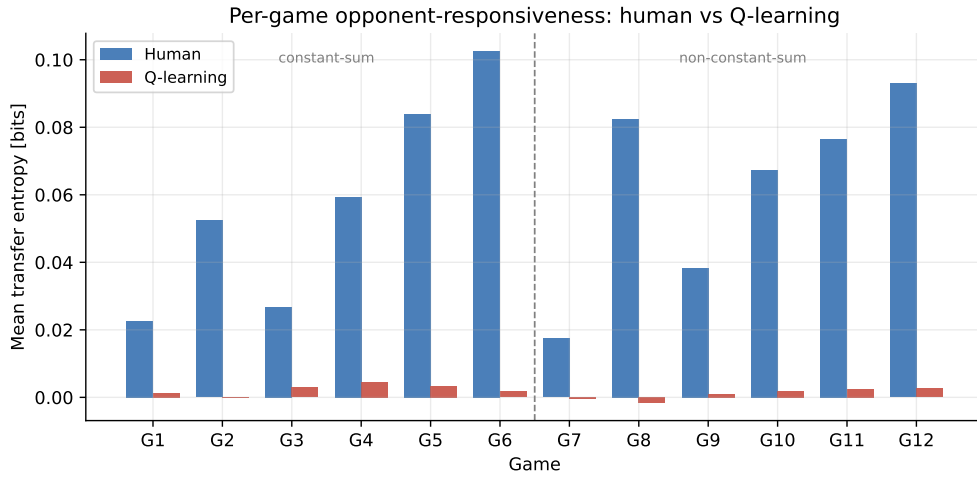


Figure 3. Per-game mean transfer entropy, humans versus matched Q-learning. The human gradient across games tracks the centrality of the mixed equilibrium; Q-learning carries no responsiveness signal in any game.

4.5 Only self-tuning EWA reaches the human region

Among the five models at their standard parameterizations, only self-tuning EWA produces human-scale transfer entropy (Table 3). The four reward-only and cumulative models sit

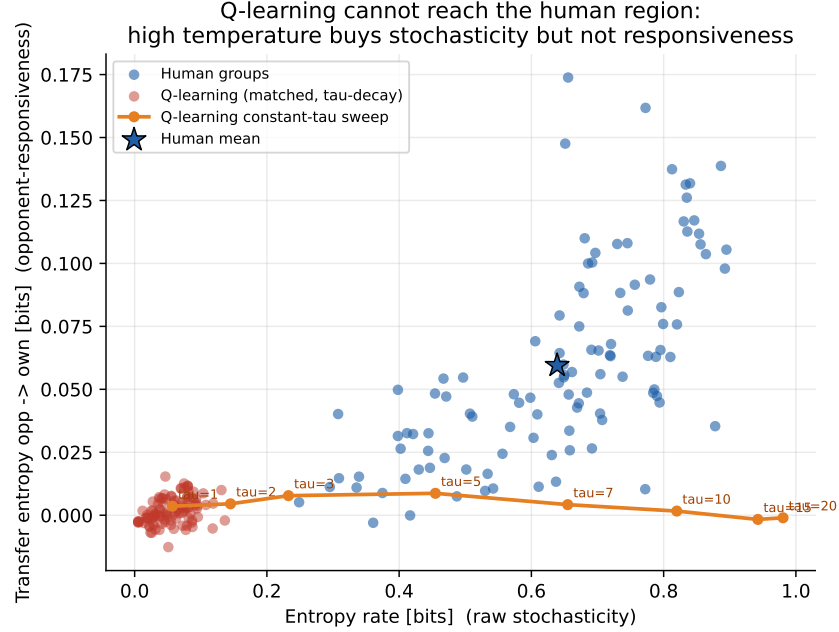


Figure 4. Transfer entropy against entropy rate. Human groups occupy the high-responsiveness region. The Q-learning temperature sweep slides along the floor: raising temperature buys stochasticity but not responsiveness.

on the transfer-entropy floor; in the (entropy rate, transfer entropy) plane, only self-tuning EWA overlaps the human cloud (Figure 5). The others fail in two distinct ways: Q-learning and reinforcement collapse to deterministic play, while fictitious play and impulse-matching produce high but information-free randomness.

Table 3. Mean signatures across all 108 matched groups, by source. Only self-tuning EWA matches human transfer entropy in order of magnitude and comes closest on entropy rate.

Source	Transfer entropy (bits)	Entropy rate (bits)
Human	0.059	0.639
Self-tuning EWA	0.115	0.707
Q-learning	0.002	0.057
Reinforcement	0.000	0.220
Fictitious play	0.001	0.789
Impulse-matching	0.002	0.922

Self-tuning EWA overshoots the human mean (0.115 against 0.059). A sensitivity sweep of its single free parameter shows the overshoot is not removable without cost: at $\lambda = 0.5$ transfer entropy matches humans (0.068) but entropy rate overshoots (0.83); at $\lambda = 1.5$ entropy rate

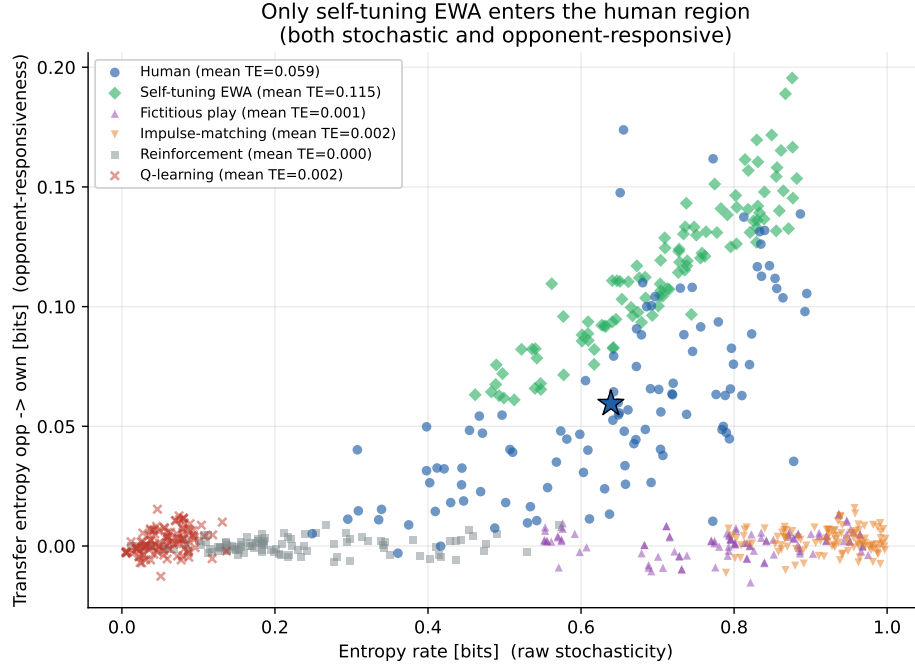


Figure 5. The (entropy rate, transfer entropy) plane for all five models and humans. Only self-tuning EWA’s cloud overlaps the human cloud. Reward-only models collapse to the lower left; cumulative belief and impulse models spread along the lower right.

matches (0.63) but transfer entropy overshoots (0.14). No single λ matches both signatures exactly, but every setting in this range keeps EWA within roughly one human standard deviation of both targets — whereas no competing model reaches the human transfer-entropy region at any examined setting.

A subject-level mixed-effects analysis confirms the contrasts with proper clustering. With a random intercept for game and 864 subjects per condition, humans differ from every model on transfer entropy: the four reward and cumulative models sit 0.058–0.061 bits below humans (all $p < 10^{-90}$), while self-tuning EWA sits 0.055 above ($p = 4 \times 10^{-72}$), the only model on the human side of zero (Figure 6; full coefficient table in Appendix C).

4.6 Robustness

Three checks address the obvious objections to the transfer-entropy result.

Estimator and history order. Table 4 recomputes the mean per-subject transfer entropy for all 864 subjects per source under three entropy estimators (plug-in, Miller–Madow, Grassberger) and two orders of own-history conditioning ($L = 1, 2$ in the generalized statistic of Appendix A). The human value is stable at 0.059–0.065 bits across all six combinations and

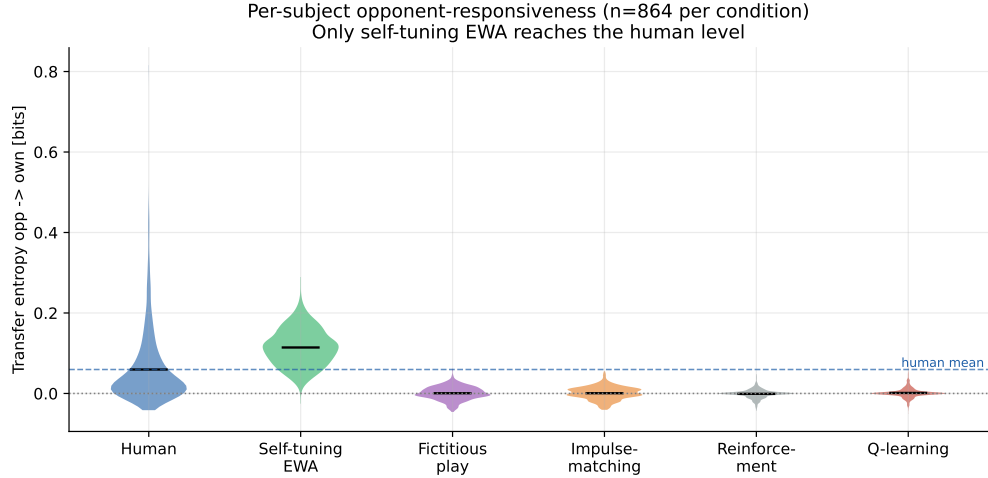


Figure 6. Per-subject transfer entropy by source ($n = 864$ per condition). Human and self-tuning EWA distributions sit above zero around the human mean; the other models collapse onto zero.

Q-learning stays at 0.001–0.002. Conditioning on a second lag of the player’s own past *raises* human transfer entropy slightly rather than removing it, so the effect is genuine opponent-responsiveness, not first-order autocorrelation leaking through the estimator.

Table 4. Robustness of mean per-subject transfer entropy (bits, shuffle-corrected; $n = 864$ subjects per source) to the entropy estimator and the order L of own-history conditioning.

Estimator	L	Human	Self-tuning EWA	Q-learning
Plug-in	1	0.059	0.114	0.002
Plug-in	2	0.064	0.127	0.002
Miller–Madow	1	0.060	0.116	0.002
Miller–Madow	2	0.065	0.133	0.002
Grassberger	1	0.059	0.114	0.001
Grassberger	2	0.065	0.131	0.002

Surrogate validation. Table 5 reports, for every individual subject, whether the raw transfer entropy exceeds the 95th percentile of a 100-permutation null in which the opponent sequence is shuffled. At a threshold calibrated to a 5% false-positive rate, 55% of human subjects pass, against 7% of Q-learning agents and 99% of self-tuning EWA agents. Human opponent-responsiveness is real at the individual level.

Window width. The windowed-entropy signatures use $w = 20$; repeating the analysis with w between 15 and 30 changes no qualitative conclusion.

Table 5. Surrogate test: share of individual subjects whose transfer entropy exceeds the 95th percentile of a 100-shuffle null ($n = 864$ per source; chance rate 5%).

Source	Share above null
Human	55.0%
Self-tuning EWA	99.1%
Q-learning	7.3%

4.7 Which architectural ingredients produce the signature

Transfer entropy requires two ingredients together: the model must attend to foregone payoffs computed against the opponent’s actual move, and it must weight recent experience enough that the last period shifts the next choice. Self-tuning EWA has both through its change-detector. Reinforcement and Q-learning lack the foregone-payoff term and never respond to the specific opponent action. Cumulative fictitious play and cumulative impulse-matching attend to the opponent but dilute the recent signal in a long-run average. Adding a recency weight ϕ to impulse-matching is consistent with this account: its transfer entropy rises from 0.005 at $\phi = 0$ (pure cumulation) to 0.034 at $\phi = 0.2$ — the human order of magnitude — and overshoots to 0.16 at $\phi = 0.5$. This matches Chmura et al. (2012), who ranked impulse-matching second to self-tuning EWA on individual round-by-round prediction; the two models that carry the recency-plus-foregone-payoff structure are exactly the two that can produce transfer entropy. We read this as identifying the architectural features associated with the signature, not as establishing that transfer entropy causes accurate choice prediction; the two are plausibly distinct measurements of a single underlying property of the learning rule.

4.8 Generality of the model contrast across game classes

The human comparison above is specific to the Selten–Chmura matching-pennies games, the only data we have. The *model-side* structural claim, however — that self-tuning EWA is simultaneously more stochastic and more opponent-responsive than Q-learning — can be tested in any game. To probe whether that contrast is peculiar to matching pennies or holds more broadly, we simulate both models under the identical protocol on seven additional canonical games spanning three classes: zero-sum (Matching Pennies, Rock–Paper–Scissors, Shapley’s game), potential (Stag Hunt, a 3×3 coordination game), and general-sum (Battle of the Sexes, Prisoner’s Dilemma). For each game we run 24 independent groups per model and compute the same shuffle-corrected transfer entropy and entropy rate, using the game’s

own action support. We emphasize that this extends only the model contrast: no human data exist for these games, so this is not a human-versus-model comparison.

Two patterns emerge (Table 6, Figure 7). The entropy-rate gap is universal: self-tuning EWA maintains higher policy entropy than matched Q-learning in all eight games, including those where both models ultimately concentrate. The transfer-entropy gap is significant (Mann–Whitney, one-sided $p < 0.05$) in six of the eight games — every game in which opponent-responsiveness is strategically relevant — and is absent in exactly the two games where it should be. In the Prisoner’s Dilemma, which is dominance-solvable, the payoff-maximizing action does not depend on the opponent, so neither model carries opponent information and both sit at the transfer-entropy floor. In Shapley’s game, a degenerate zero-sum cyclus in which learning dynamics famously fail to settle, both models sit near the floor and are statistically indistinguishable. The contrast therefore generalizes across game classes wherever responsiveness can matter, and vanishes, as it should, where it cannot. This is a model-level robustness check, not a claim about human play in these games; establishing that humans exhibit the same signatures outside matching pennies would require new experimental data and is the natural next step (Section 5).

Table 6. Model-side signatures across game classes (24 groups per model per game; mean $\pm$ SD for transfer entropy). The entropy-rate gap (EWA $>$ Q-learning) holds in every game; the transfer-entropy gap is significant wherever opponent-responsiveness is strategically relevant, and absent in the dominance-solvable Prisoner’s Dilemma and the degenerate Shapley cyclus. p is a one-sided Mann–Whitney test that EWA transfer entropy exceeds Q-learning’s.

Game	Class	EWA TE	Q-learn TE	p	ER gap
SC matching pennies	zero-sum (data)	0.137 ± 0.027	0.003 ± 0.005	1.5×10^{-9}	+0.75
Matching Pennies	zero-sum	0.024 ± 0.012	0.006 ± 0.008	3.1×10^{-8}	+0.09
Rock–Paper–Scissors	zero-sum	0.027 ± 0.024	0.003 ± 0.011	2.2×10^{-6}	+0.12
Shapley	zero-sum	0.018 ± 0.018	0.025 ± 0.028	0.89	+0.37
Stag Hunt	potential	0.088 ± 0.085	0.010 ± 0.018	6.2×10^{-8}	+0.43
Coordination 3×3	potential	0.079 ± 0.038	0.048 ± 0.024	2.7×10^{-6}	+0.86
Battle of the Sexes	general-sum	0.084 ± 0.026	0.019 ± 0.024	1.5×10^{-9}	+0.71
Prisoner’s Dilemma	general-sum	0.007 ± 0.009	0.010 ± 0.018	0.72	+0.45

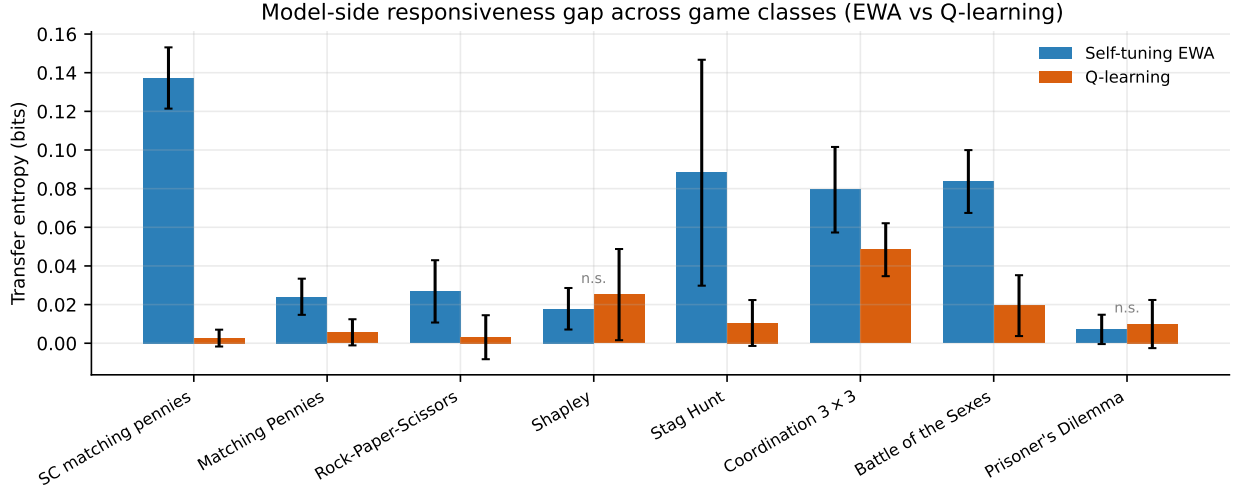


Figure 7. Transfer entropy of self-tuning EWA versus matched Q-learning across game classes (24 groups per model per game; error bars are ± 1 SD). EWA carries more opponent-to-self information in every game where responsiveness is strategically relevant; the two cases marked “n.s.” (Shapley, Prisoner’s Dilemma) are the degenerate cycler and the dominance-solvable game, where both models sit at the transfer-entropy floor. No human data are involved in this comparison.

5. Discussion

Human play in these games is noisy and adaptive at once. Subjects maintain a high, roughly stationary level of policy entropy and, within that noise, condition their next move on the opponent’s last move — the more so the closer the game’s equilibrium lies to the center of the simplex. The dominant computational model of learning, reinforcement with decaying exploration, reproduces neither feature and cannot be tuned to reproduce both, because raising its stochasticity destroys its responsiveness. Among standard learning models, only self-tuning EWA occupies the human region of the signature plane.

This gives an information-theoretic reading of the model-selection result of Chmura et al. (2012). Their criterion was choice prediction: how often a model calls the next action correctly. Ours is a property of the trajectory: how much information flows from opponent to self, and how stochastic play stays. The two criteria are unrelated, and they select the same model. The transfer-entropy signature is consistent with this convergence: self-tuning EWA is the only standard learner that is simultaneously stochastic and opponent-responsive, the structure human play demands. We do not claim that carrying transfer entropy *causes* accurate choice prediction; more cautiously, both are properties of a learning rule that couples recency-weighted updating to foregone-payoff attention, and EWA is the one standard rule that has both. It also clarifies the relationship to the stationary concepts that perform well

on this data’s aggregate frequencies Selten and Chmura (2008); Brunner et al. (2011): a concept like impulse balance equilibrium describes where play settles, while the trajectory signatures describe the adjustment process that holds it there, and the adjustment process associated with human-like signatures is the recency-weighted, foregone-payoff-driven one.

The signatures are model-free and transfer to any repeated-game dataset with period-level actions. They offer a way to screen learning models before estimation: a model that cannot match a dataset’s transfer entropy and entropy rate need not be fit, whatever its choice-prediction score on aggregate frequencies. The cross-game-class analysis of Section 4.8 shows that the discriminating contrast — EWA simultaneously stochastic and responsive where Q-learning cannot be — is not an artifact of matching pennies: it recurs across zero-sum, potential, and general-sum games, and recedes exactly where opponent-responsiveness stops being strategically relevant.

Several limitations bound the claims, and we frame them as the agenda this paper opens rather than as caveats alone. First, the contribution is methodological: we introduce a diagnostic dimension — which dynamical signatures a learning model must reproduce — rather than a new learning rule or equilibrium concept, and we apply it to an existing benchmark dataset with existing models. The value is in the screen and the mechanism it exposes, not in a new model; building a learning rule designed to match the human signatures directly is a natural follow-up. Second, the human evidence comes from one game class — 2×2 generalized matching pennies with completely mixed equilibria — and binary actions. The model-side contrast generalizes across game classes (Section 4.8), but the decisive test — whether *human* subjects display the same maintained-entropy, opponent-responsive signatures in richer games (Rock–Paper–Scissors, congestion and public-goods games, repeated Prisoner’s Dilemma, auctions, larger action sets) — requires new experimental data and is the single most valuable extension; our protocol and estimators are designed to apply unchanged once such data exist. Third, the relationship between the transfer-entropy signature and choice-prediction skill is associational: both are properties of a recency-weighted, foregone-payoff-driven learning rule, and we do not claim one causes the other. Fourth, the models are compared at standard parameterizations under a screening logic, not estimated subject by subject; the sensitivity analyses cover the dimensions that matter for our conclusions (exploration temperature, EWA precision, impulse recency) but not the full joint parameter space. Within those bounds the central finding is robust: human play carries directed information from opponent to self at a level no reward-averaging or cumulative learner reproduces, and the one standard model that does carry it is the one that best predicts human choices.

Data and code availability. All analyses are produced by a single self-contained Python codebase from fixed random seeds; it implements the estimators, the five learning models, the simulation protocol, and every analysis driver. The code, the stored numerical outputs, a reproducibility guide, and the full per-game and cross-game tables accompany this submission as Supplementary Material. The period-by-period Selten–Chmura data were obtained from the original authors and are not redistributed.

Funding. This research received no external funding.

A. Signature estimators

Policy-entropy trajectory

For a player with binary action sequence $a_1, \dots, a_T$ we estimate the mixing probability in a sliding window covering periods t through $t + w - 1$, with $w = 20$,

$$\hat{p}_t = \frac{1}{w} \sum_{s=t}^{t+w-1} \mathbf{1}[a_s = 0], \quad H_t = -\hat{p}_t \log_2 \hat{p}_t - (1 - \hat{p}_t) \log_2 (1 - \hat{p}_t),$$

giving a trajectory of length $T - w + 1$. For the group-level oscillation and KL-fluctuation signatures we pool the four same-role subjects within the window before forming $\hat{p}_t$; for the entropy-decline signature we form H_t per subject and average, so the decline measures the annealing of an individual’s own policy. Results are insensitive to w in the range 15–30.

The *oscillation amplitude* is the peak magnitude of the one-sided discrete Fourier transform of the linearly detrended H_t , normalized by series length and excluding the zero frequency. The *KL fluctuation* is the standard deviation of the detrended series $1 - H_t$, since for a binary policy $\text{KL}(\pi_t \parallel \text{uniform}) = \log_2 2 - H_t = 1 - H_t$. The *entropy decline* is the ordinary-least-squares slope of H_t on t .

Transfer entropy

For source sequence x (the opponent action faced) and target y (the player’s own action), with support sizes k_x, k_y and target history order L , define the composite history state $y_t^{(L)} = (y_{t-1}, \dots, y_{t-L})$. The transfer entropy is

$$T_{x \rightarrow y}^{(L)} = [H(y_t, y_t^{(L)}) - H(y_t^{(L)})] - [H(y_t, y_t^{(L)}, x_{t-1}) - H(y_t^{(L)}, x_{t-1})],$$

the reduction in conditional entropy of y_t from adding x_{t-1} to the target’s own L -lag past Schreiber (2000). We report $L = 1$ in the main text and confirm robustness at $L = 2$

(Table 4). Each joint entropy is estimated with the Grassberger finite-sample correction Grassberger (2008). To remove residual bias we subtract a shuffle baseline in the spirit of the effective transfer entropy of Marschinski and Kantz (2002): the mean transfer entropy over 50 random permutations of x , which destroys the coupling while preserving marginals. The surrogate test of Table 5 instead compares the raw statistic to the 95th percentile of a 100-permutation null.

Entropy rate

The reported entropy rate is the order-1 Markov conditional entropy $H(a_t \mid a_{t-1})$ of a subject’s action sequence, estimated as $H(a_{t-1}, a_t) - H(a_{t-1})$ with the Grassberger correction. It is the single-number summary of how stochastic play remains.

B. Learning models

Each agent holds its own payoff matrix M indexed $M(a, o)$ by own action a and opponent action o (row agents use the game’s row-payoff matrix, column agents the transpose of the column-payoff matrix). All models are simulated under the experimental protocol of Section 3.3; each agent updates from the action of the opponent it actually faced.

Q-learning. Boltzmann exploration over action values,

$$Q(a) \leftarrow Q(a) + \alpha[r + \gamma \max_b Q(b) - Q(a)], \quad \pi_t(a) = \frac{\exp(Q(a)/\tau_t)}{\sum_b \exp(Q(b)/\tau_t)}, \quad \tau_t = \tau_0 \rho^t,$$

with $\alpha = 0.1$, $\gamma = 0.95$, $\tau_0 = 1$, and decay $\rho = 0.99$ for the matched comparison; the temperature sweep in Section 4.4 varies the constant temperature τ with $\rho = 1$.

Reinforcement learning Erev and Roth (1998). Propensities reinforced by the realized payoff of the chosen action,

$$q(a) \leftarrow q(a) + r \mathbf{1}[a = a_t], \quad \pi(a) = \frac{q(a)}{\sum_b q(b)}.$$

Stochastic fictitious play Fudenberg and Levine (1998). A logit best response to the running empirical frequency of opponent play,

$$c(o) \leftarrow c(o) + \mathbf{1}[o = o_t], \quad \pi(a) \propto \exp\left(\beta \sum_o \frac{c(o)}{\sum_{o'} c(o')} M(a, o)\right),$$

with $\beta = 1$.

Self-tuning EWA Ho et al. (2007). Attractions $A(a)$ and an experience weight N updated with a self-tuning change detector and an attention function. Let $h^o(t)$ be the cumulative frequency of opponent action o through period t and $r^o(t) = \mathbf{1}[o_t = o]$ the immediate indicator. The change detector and attention are

$$\phi(t) = 1 - \frac{1}{2} \sum_o (h^o(t) - r^o(t))^2, \quad \delta^a(t) = \mathbf{1}[M(a, o_t) \geq M(a_t, o_t)].$$

With $\kappa = 0$,

$$N(t) = \phi(t) N(t-1) + 1,$$

$$A^a(t) = \frac{\phi(t) N(t-1) A^a(t-1) + [\mathbf{1}[a=a_t] + \delta^a(t) \mathbf{1}[a \neq a_t]] M(a, o_t)}{N(t)},$$

and $\pi(a) \propto \exp(\lambda A^a)$ with $\lambda = 1$. The detector $\phi(t)$ falls when recent opponent play departs from its history, down-weighting stale experience; $\delta^a(t)$ directs attention to foregone payoffs that would have done at least as well.

Impulse-matching Chmura et al. (2012). Following impulse balance theory, losses relative to the agent's pure-strategy maximin (security level) $s = \max_a \min_o M(a, o)$ are counted double. The transformed payoff matrix is

$$\tilde{M}(a, o) = \begin{cases} M(a, o), & M(a, o) \geq s, \\ 2 M(a, o) - s, & M(a, o) < s, \end{cases}$$

which doubles the shortfall of any payoff below the security level. After each round the agent forms an impulse toward any action that would have paid more in the transformed game against the realized opponent action, and chooses with probability proportional to cumulated impulse:

$$q(a) \leftarrow (1 - \phi) q(a) + \max(0, \tilde{M}(a, o_t) - \tilde{M}(a_t, o_t)), \quad \pi(a) = \frac{q(a)}{\sum_b q(b)}.$$

The canonical model is the pure-cumulation case $\phi = 0$, used in all headline comparisons; the mechanism analysis in Section 4.7 varies the recency weight ϕ .

C. Mixed-effects estimates

Table 7 reports the fixed-effect contrast of each model against humans from linear mixed-effects models on per-subject signatures with a random intercept for game (864 subjects per condition; the model simulations here are an independent replication of the group-level runs, drawn with different random seeds, which is why the model means differ from Table 3 within sampling error).

Table 7. Mixed-effects contrasts, model minus human, per-subject signatures, random intercept for game. Human means: transfer entropy 0.060 (SD 0.086), entropy rate 0.639 (SD 0.305).

Signature	Model	Difference	95% CI	p
Transfer entropy	Self-tuning EWA	+0.055	[+0.049, +0.061]	4×10^{-72}
	Q-learning	-0.058	[-0.064, -0.053]	2×10^{-91}
	Reinforcement	-0.060	[-0.066, -0.055]	2×10^{-98}
	Fictitious play	-0.059	[-0.065, -0.053]	7×10^{-92}
	Impulse-matching	-0.059	[-0.065, -0.053]	2×10^{-91}
Entropy rate	Self-tuning EWA	+0.073	[+0.053, +0.093]	4×10^{-13}
	Q-learning	-0.588	[-0.608, -0.569]	$< 10^{-300}$
	Reinforcement	-0.357	[-0.387, -0.328]	5×10^{-122}
	Fictitious play	+0.148	[+0.125, +0.171]	1×10^{-37}
	Impulse-matching	+0.285	[+0.265, +0.304]	2×10^{-180}

References

- Brunner, C., Camerer, C.F., Goeree, J.K.: Stationary concepts for experimental 2×2 -games: Comment. *Am. Econ. Rev.* **101**(2), 1029–1040 (2011)
- Camerer, C., Ho, T.-H.: Experience-weighted attraction learning in normal form games. *Econometrica* **67**(4), 827–874 (1999)
- Chmura, T., Goerg, S.J., Selten, R.: Learning in experimental 2×2 games. *Games Econ. Behav.* **76**(1), 44–73 (2012)
- Erev, I., Roth, A.E.: Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. *Am. Econ. Rev.* **88**(4), 848–881 (1998)
- Fudenberg, D., Levine, D.K.: *The Theory of Learning in Games*. MIT Press, Cambridge, MA (1998)
- Grassberger, P.: Entropy estimates from insufficient samplings. *arXiv:physics/0307138* (2008)
- Ho, T.-H., Camerer, C.F., Chong, J.-K.: Self-tuning experience weighted attraction learning in games. *J. Econ. Theory* **133**(1), 177–198 (2007)

- Author: Decoding adaptive noise: information-theoretic signatures of learning dynamics in repeated games. [Anonymized for double-blind review; full citation to be provided upon acceptance] (2025)
- Marschinski, R., Kantz, H.: Analysing the information flow between financial time series: An improved estimator for transfer entropy. *Eur. Phys. J. B* **30**(2), 275–281 (2002)
- McKelvey, R.D., Palfrey, T.R.: Quantal response equilibria for normal form games. *Games Econ. Behav.* **10**(1), 6–38 (1995)
- Schreiber, T.: Measuring information transfer. *Phys. Rev. Lett.* **85**(2), 461–464 (2000)
- Selten, R., Buchta, J.: Experimental sealed bid first price auctions with directly observed bid functions. In: Budescu, D., Erev, I., Zwick, R. (eds.) *Games and Human Behavior*. Lawrence Erlbaum (1999)
- Selten, R., Chmura, T.: Stationary concepts for experimental 2×2 -games. *Am. Econ. Rev.* **98**(3), 938–966 (2008)
- Selten, R., Chmura, T., Goerg, S.J.: Stationary concepts for experimental 2×2 -games: Reply. *Am. Econ. Rev.* **101**(2), 1041–1044 (2011)

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [supplementary.pdf](#)