Heads or Tails?
A Bayesian Framework for Bias Detection in Sequence-Dependent Betting

Vojtěch Zíka

January 26, 2026

Abstract

This project investigates how individuals interpret random sequences and translate these interpretations into risky betting decisions. In a controlled laboratory experiment, participants repeatedly decide whether to place a bet, how much to stake, and which side to bet on in a fair coin toss after observing a history of six prior outcomes. Building on formal predictions of expected utility theory, the design features two treatments: one in which betting is normatively optimal for all but extremely risk-averse participants, and one in which abstention is optimal for all but extremely risk-loving participants. Using individual risk parameters estimated from the Holt–Laury task, Bayesian hierarchical models assess whether observed betting behavior is normatively optimal and, when it is not, quantify systematic over- or underbetting and the associated expected utility losses. The analysis further identifies which specific sequences trigger these deviations. Beyond the extensive and intensive margins of betting, we examine directional side choices to determine whether participants exhibit hot-hand or gambler's-fallacy tendencies. These tendencies are inferred empirically using a Bayesian anchor-based framework that maps each sequence to diagnostic pure streaks, rather than being imposed ex ante. Finally, we examine how dispositional optimism and response times relate to hot-handish versus gamblerish responding and assess the correspondence between anchor-based and canonical heuristic classifications. To complement the laboratory evidence, the study also includes a small classroom-based sanity-check experiment using physically realized coin tosses. This auxiliary study is explicitly non-confirmatory and provides a descriptive benchmark for comparing behavior across experimental settings, while all confirmatory inference relies on the preregistered laboratory experiment.

Contents

1 Study Overview **4**

1.1 Main Laboratory Study . . . 4

1.1.1 Experimental Design . . . 4

1.1.2 Participants and Procedure . . . 6

1.1.3 Pilot Testing, Verification, and Preregistration Updates . . . 6

1.2 Sanity-Check Classroom Study . . . 7

1.2.1 Experimental Design . . . 7

1.2.2 Participants and Procedure . . . 8

1.3 Materials and Code Availability . . . 9

2 Statistical Analysis **9**

2.1 Confirmatory Analyses . . . 10

2.1.1 RQ1: Do any sequences systematically affect the likelihood of betting relative to the expected-utility–optimal choice? . . . 10

2.1.2 RQ2: Do any sequences lead to over- or under-staking relative to the expected-utility–optimal amount? . . . 11

2.1.3 RQ3: Do any sequences increase certainty-equivalent losses from suboptimal betting? . . . 13

2.1.4 RQ4: Do any sequences systematically bias side choices toward Heads or Tails? . . . 14

2.2 Exploratory Analyses . . . 16

2.2.1 EX1: Hot-handish and gamblerish patterns in side choice . . . 16

2.2.2 EX2: Participant-level associations with optimism and response time across RQ1–RQ4 . . . 19

2.2.3 EX3: False–Positive treatment replication and FN–FP contrasts . . . 20

2.2.4 EX4: Similarity Between Anchor-Based and Canonical Rule-Based Classifications . . . 21

A Bayesian Modelling Framework and Robustness Procedures **24**

A.1 Priors . . . 24

A.2 Likelihoods . . . 24

A.3 Posteriors . . . 24

A.4 Session Drift Adjustment . . . 24

A.5 Sensitivity and Robustness Analyses . . . 24

A.6 Parameters and Indices . . . 25

A.6.1 Holt–Laury Risk Preferences (r) . . . 25

A.6.2 Expected-Utility–Optimal Stakes (a, a^*) . . . 25

A.6.3 Certainty Equivalent (c) 26

B Experimental Materials **27**

B.1 Sequence Betting Task Decision Interface 27

1 Study Overview

Understanding how individuals interpret random sequences and translate these interpretations into risky choices is central to models of decision-making under risk. This study employs an incentive-compatible experimental design that provides a clean environment for measuring sequence-level distortions and individual heterogeneity relative to normative expected-utility (EU) benchmarks. In each trial, participants observe a history of six coin flips and then decide whether to place a bet, how much to stake, and which side (Heads or Tails) to bet on for the subsequent flip.

The False-Negative (FN) treatment uses a favorable return multiplier ($m = 2.5$), for which betting is EU-optimal in all sequences for all but extremely risk-averse participants. This treatment provides a well-defined benchmark for assessing underbetting, stake miscalibration, and welfare losses, and therefore forms the basis for all confirmatory analyses. The False-Positive (FP) treatment uses an unfavorable return multiplier ($m = 1.9$), under which the EU-optimal stake is zero for all but extremely risk-loving participants. Because deviations in this setting are mechanically one-sided and the intensive margin is degenerate, FP serves primarily as an exploratory and robustness condition.

This design also supports an additional exploratory layer of analysis. Within the Bayesian modelling framework, hot-handish (streak–following) and gamblerish (streak–reversing) tendencies are inferred directly from observed choices, without relying on any *ex ante* assumptions about what constitutes a streak. Subsequent analyses then examine whether these tendencies covary with individual characteristics—including dispositional optimism, risk preferences, and response time—thereby providing insight into the behavioral mechanisms that may drive deviations from EU–optimal behavior.

In addition to the main laboratory experiment (Section 1.1), the study includes a small classroom-based sanity-check experiment (Section 1.2), designed to provide a descriptive benchmark using physically realized coin tosses. The main laboratory study relies on algorithmically generated sequences in order to ensure full coverage of the six-toss sequence space, statistical independence across trials, and sufficient sample size for sequence-level inference; the classroom study serves only as an auxiliary, explicitly non-confirmatory comparison and is analyzed separately. This auxiliary classroom study is intended to acknowledge concerns about external validity that arise from the tightly structured and highly controlled nature of the main laboratory experiment, while preserving the main study's ability to deliver clean sequence-level inference.

1.1 Main Laboratory Study

1.1.1 Experimental Design

The experiment elicits betting decisions on a fair coin toss conditional on observing a history of six prior outcomes. In each round, participants are shown a hypothetical six-toss sequence representing a possible realization of independent fair coin tosses; these sequences do not correspond to outcomes from earlier rounds, and participants are explicitly informed of this fact. Six-toss sequences are the shortest even-length histories that generate a sufficiently rich sequence space while including diagnostic patterns such as `HHHTTT` and `TTTHHH`; shorter histories are too limited, whereas longer histories would substantially expand the sequence space and dilute per-sequence information. Rather than relying on randomly realized histories, the design deliberately presents all sequences in a balanced manner and in random order to ensure symmetric and complementary trials (e.g., Heads–Tails reversals), which permits clean theoretical benchmarks and sequence-level inference under the assumption of a fair coin.

Each participant completes 64 decision rounds, one for each unique six-toss sequence (e.g., `HHHHHH`, `THHHHH`, ...). In every round, participants decide whether and how to bet on the coin toss that would follow the displayed history. If they choose to bet, they select a side (Heads or Tails) and a stake from a fresh 100-ECU endowment. Stakes are integers from 1 to 100 ECU; choosing "No bet" corresponds to a stake of 0.

Payoffs are incentive-compatible. If the coin toss following the displayed sequence matches the chosen side, the participant earns $e - b + mb$; otherwise, the payoff is $e - b$, where $e = 100$ ECU is the endowment, b is the stake, and m is the payout multiplier. The multiplier is manipulated between subjects.

- False-Negative (FN) treatment, $m = 2.5$. Under expected-utility theory, all but extremely risk-averse participants should place a positive bet; deviations primarily take the form of false negatives (not betting when betting is optimal). The value $m = 2.5$ was chosen as a compromise between providing a sufficiently strong incentive to bet for most participants and avoiding corner solutions in which participants always bet their full endowment.
- False-Positive (FP) treatment, $m = 1.9$. Expected-utility theory predicts that all but sufficiently risk-loving participants should refrain from betting; deviations therefore mainly consist of false positives (betting when abstention is optimal). The multiplier $m = 1.9$ implies an expected value comparable to outside bets in American roulette with two zeros.

The decision interface displays the six-outcome history, options for *No bet*, *Heads*, and *Tails*, a stake slider, and a payoff preview based on the current stake. Each round starts with a fresh 100-ECU endowment. At the end of the experiment, one round is randomly selected for payment and converted to cash at a posted exchange rate. The decision interface presented to participants is shown in Appendix B.1.

Before the main task, participants complete standardized instructions, a short comprehension quiz, and one practice round. After completing the 64 decision rounds, they perform the Holt–Laury risk elicitation task (to estimate CRRA risk preferences) and complete the LOT-R questionnaire. Response time from screen onset to choice confirmation is recorded for descriptive analyses.

This within-subject sequence design, combined with between-subject variation in the incentive multiplier, enables clean identification of bias-driven deviations from expected-utility benchmarks while holding objective risk and return parameters constant.

Randomization is implemented at multiple levels to ensure balanced and unbiased assignment. All randomization procedures are implemented in oTree using its default pseudorandom number generator, ensuring reproducible and unbiased draws across participants and sessions.

- PC terminal randomization. Before each session, participants draw seat numbers at random to assign them to laboratory terminals, ensuring that treatment assignment and sequence presentation are independent of seating order or social proximity.
- Treatment randomization. Participants are randomly assigned to either the FP or FN treatment at session start. Treatment groups are expected to be unbalanced, approximately in a 3:1 ratio favoring the FN treatment, since fewer bets are expected in the FP condition and a larger FN sample is required for stable estimation.
- Sequence randomization. Each participant sees all 64 six-toss sequences once, presented in a unique random order and partitioned into four equal blocks of 16 sequences to enable drift checks.

- Seventh-toss randomization. The coin outcome determining payoffs is generated independently for each trial, with an equal 50% probability of Heads or Tails.
- Interface randomization. In each decision round, the positions of the betting buttons for *Heads* (Hlava) and *Tails* (Orel) are randomized between the first two positions on the screen. The *No bet* (Nesázet) option is always displayed in a fixed third position. This design prevents systematic left–right response biases in side choice while preserving a stable and salient location for the abstention option, which is critical for identifying false negatives and false positives on the extensive margin of betting.

1.1.2 Participants and Procedure

Participants will be recruited from the ORSEE database maintained by the Laboratories of Behavioral Studies at Jan Evangelista Purkyně University in Ústí nad Labem. The database currently includes approximately 1,400 registered individuals, consisting primarily of university students. The sex composition of the sample is expected to reflect the underlying ORSEE subject pool, which is approximately 2:1 female to male; no stratification by sex is implemented. Participation will be voluntary and compensated with a show-up fee plus an additional performance-based payment.

The final sample size will be limited by the recruitment capacity of the database. We expect to collect data from approximately 200–250 participants in total, randomly assigned to the FP and FN treatments in an expected ratio of about 1:3. The experimental design will be implemented in oTree, and data collection is planned for February 2026.

No deception will be used in the experiment, and all procedures will comply with the ethical standards of Jan Evangelista Purkyně University and the guidelines for behavioral and experimental research established by the Laboratories of Behavioral Studies.

Although the study is conducted with laboratory participants, the task closely mirrors real-world probabilistic forecasting and speculative investment under repeated exposure to random outcomes. With an expected sample of 200–250 participants (yielding roughly 12,800–16,000 sequence-level observations), the design provides sufficient precision for hierarchical Bayesian estimation of both sequence- and participant-level effects. This enables reliable inference on small behavioral deviations and allows the derived behavioral parameters to inform structural models of sequential belief updating and risk-taking.

No participants will be excluded based on behavioral outcomes. Data-quality exclusions, if any, will be limited to technical failures (e.g., incomplete sessions, missing variables) or violations of task execution (e.g., failure to complete required stages). Response-time extremes and dominant choice patterns are accommodated by the hierarchical models rather than excluded.

1.1.3 Pilot Testing, Verification, and Preregistration Updates

This preregistration is finalized prior to any pilot data collection that could inform modeling or estimand choices. The preregistration will be uploaded and time-stamped prior to conducting any pilot sessions. The pilot study will then be used exclusively to verify the executability of the preregistered design and analysis plan. Specifically, pilot data will be inspected only to assess: (i) whether all preregistered variables are correctly recorded and stored; (ii) whether randomization, payoff calculations, and task logic function as specified; (iii) whether the data structure supports the construction of all preregistered estimands and derived quantities; and (iv) whether the preregistered Bayesian models are computationally well defined and numerically stable when applied to real data.

Pilot data from the behavioral task will not be used to evaluate behavioral patterns, effect sizes, or the plausibility of hypotheses. No decisions about hypotheses, estimands, outcome definitions, or inferential criteria will be based on observed pilot outcomes. Any inspection of pilot behavioral data is strictly limited to implementation, consistency, and feasibility checks that could, in principle, also be identified using simulated data.

If the pilot reveals issues that prevent faithful execution of the preregistered analysis plan (e.g., missing variables, incorrect data recording, misaligned scales, or computational instability of preregistered models), a preregistration addendum will be prepared and uploaded prior to the start of formal data collection. The addendum will be time-stamped, explicitly linked to the original preregistration, and will document all changes relative to the original preregistration together with a justification for each change. All such changes will be limited to technical or procedural refinements. Substantive research questions, estimands, hypotheses, and inferential criteria will not be altered based on pilot data.

Following completion and upload of any such addendum, the preregistration will be considered final and locked. Formal data collection for the main study will begin only after this point. Pilot behavioral observations will be excluded from all confirmatory and exploratory analyses reported in this study.

In the pilot phase only, eye-tracking data may be collected as an auxiliary process measure. These data may be used to support user-interface verification and, where informative, to provide descriptive illustrations of how participants allocate visual attention during decision making (e.g., attention to the sequence display, payoff preview, or choice controls). Eye-tracking data are not part of the preregistered outcomes and will not be used for hypothesis testing, model estimation, or confirmatory inference. Any use of eye-tracking data will be limited to qualitative or descriptive interpretation in the Results or Discussion sections.

1.2 Sanity-Check Classroom Study

In addition to the main preregistered laboratory experiment, a small classroom-based sanity-check study will be conducted using a convenience sample of approximately 20–30 students. This study is intended solely to benchmark whether qualitative sequence-dependent betting patterns observed in the main experiment appear similar when the payoff-relevant outcome is generated by a physically tossed coin in real time. Owing to the small sample size and the limited number of realized seventh-toss outcomes per sequence, analyses will be restricted to descriptive and qualitative summaries (e.g., betting rates, stake distributions, and directional side choices conditional on betting) and will not involve hierarchical modeling, hypothesis testing, or confirmatory inference, nor will they be used to update preregistered conclusions.

1.2.1 Experimental Design

The classroom task mirrors the main study in its core decision structure but differs in how sequences are generated. Participants face 64 six-toss sequences that are realized in real time by repeated physical coin tossing. The 64 sequences are collected sequentially and divided into four blocks of 16 realized sequences. Because sequences are generated by chance, some six-toss sequences may occur multiple times, while others may not occur at all. For each realized sequence, participants decide whether to place a bet, which side to bet on, and how much to stake. As in the main study, one decision is randomly selected for payment at the end of the task. Given the small classroom sample size, the classroom study will use a single favorable payout multiplier ($m = 2.5$; FN frame). The unfavorable multiplier used in the main study will not be implemented in the classroom study.

After the classroom task, participants complete a short Holt–Laury risk task and the LOT-R. In the classroom study, these measures will be reported descriptively (e.g., means/SDs and distributions) to benchmark the classroom sample against the main study sample; they will not be used for inferential modeling in the classroom study.

1.2.2 Participants and Procedure

Participants will be students enrolled in a course taught by the principal investigator at the Faculty of Social Sciences, Charles University. The classroom study will be conducted in the second week of the semester (end of February 2026), prior to any graded coursework, to minimize potential dependence or perceived pressure. Participation is voluntary, and participation or non-participation has no effect on course evaluation. Participants will receive monetary compensation based on their experimental decisions, independent of course requirements.

At the beginning of the session, two participants are randomly selected to serve distinct roles: a coin tosser and a recorder. Random selection is implemented by having each participant draw one folded slip from an opaque container containing $N-2$ blank slips, one slip marked "TOSSER," and one slip marked "RECORDER." The participant assigned the role of coin tosser performs all coin tosses using a standard coin. The coin used for all tosses is taken from the personal belongings of either the tosser or the recorder. If neither the tosser nor the recorder has a coin available, participants who have a coin place it visibly on their desk. The coin used is the one physically closest to the whiteboard, as determined by visual inspection. Only if no participant has a coin available is a coin provided by the experimenter, and this is explicitly disclosed to all participants. The participant assigned the role of recorder privately observes and records the realized outcomes and writes the current six-toss sequence on the whiteboard for participants to observe. After the completion of the second block (i.e., after 32 realized sequences), the two selected participants switch roles for the remaining blocks. All role assignments and role changes are conducted publicly.

For each of the 64 sequences, a seventh coin toss is generated by the selected tosser. To prevent learning or belief updating from realized outcomes during the task, the outcome of each seventh toss is generated but not revealed to decision makers until all 64 decisions have been recorded. The toss is performed into an opaque cup or under an opaque cover. A designated recorder privately observes and records the realized outcome for each sequence on a pre-printed outcome log that is not visible to other participants.

Each sequence is presented separately. For a given sequence, the recorder publicly posts the outcomes of six consecutive coin tosses on the whiteboard using cards labeled H or O, such that only the current six-toss sequence is visible at any time. After the sixth toss is posted, participants are given a fixed decision interval to record their choice on a decision sheet, which records only the trial number and the participant's decision, but not the sequence itself. After the decision interval ends, all cards are removed from the board before the next sequence begins. This procedure ensures that participants never observe more than one six-toss sequence at a time and do not have access to a cumulative history of sequences.

After all 64 decisions have been completed, one trial is selected at random for payment using a transparent physical randomization device (e.g., drawing one numbered slip from an opaque container containing numbers 1–64). Participants are paid according to the same payoff function as in the main study, using the previously recorded (and unrevealed) outcome of the seventh toss corresponding to the selected sequence.

1.3 Materials and Code Availability

All preregistration materials are hosted on the project's OSF repository. Following data collection and anonymization, the OSF repository will contain the experimental data, preregistered analysis scripts, and documentation necessary to reproduce all confirmatory and exploratory analyses reported in this study.

OSF repository
`https://osf.io/v5hzs/overview?view_only=fe797f270f2e4214a74c3cf622d9ca49`

The experimental task for the main laboratory study was implemented in oTree. The corresponding oTree source code is available in a GitHub repository, which documents task logic, randomization, and payoff implementation. The GitHub repository is provided for transparency and reproducibility and does not substitute for the preregistered design and analysis plan described in this document.

GitHub repository
`https://github.com/vojtechzika/sequence-betting`

2 Statistical Analysis

All preregistered statistical analyses are conducted within a unified Bayesian modelling framework and apply exclusively to the main laboratory study (Section 1.1). General estimation conventions, prior specifications, likelihood formulations, and robustness procedures applicable to these models are described in Appendix A. The classroom-based sanity-check study (Section 1.2) is analyzed descriptively and separately, as specified in its corresponding section.

Confirmatory analyses are conducted in the False–Negative treatment, where betting is expected-utility–optimal for the vast majority of participants due to the favorable return multiplier $m = 2.5$. This provides a single normative benchmark for evaluating (RQ1) the extensive margin of betting (bet vs. no bet), (RQ2) calibration of stake size relative to the EU–optimal benchmark, (RQ3) money–metric welfare loss (certainty–equivalent difference), and (RQ4) sequence–induced directional bias in side choice. Definitions of all derived quantities—risk parameter r, optimal stakes a^*, certainty equivalents c—are provided in Appendix A.6. Confirmatory evidence is evaluated using directional posterior probabilities rather than classical null–hypothesis tests, with the Bayesian expectations and evidence criteria specified separately for each RQ. Throughout the analysis, practical equivalence thresholds (e.g., ρ, δ) are used solely to summarize posterior concentration and to facilitate exposition. These thresholds do not define acceptance or rejection of hypotheses; all substantive conclusions are based on full posterior distributions and reported uncertainty.

Exploratory analyses extend the framework in four directions. EX1 constructs a unified Gamblerish–Hot-handish Index (GHI) from anchor-based sequence similarity, tests whether individual characteristics predict participant-level GHI, and assesses whether responses to the pure sequences provide a sufficient approximation to the full index. EX2 examines how dispositional optimism and response time relate to the confirmatory behavioral measures from RQ1–4. EX3 provides descriptive contrasts for RQ1 and RQ3 between the False–Negative treatment and the False–Positive treatment, where non-betting is EU–optimal for all but extremely risk–loving participants due to the unfavorable return multiplier $m = 1.9$. EX4 evaluates the correspondence between anchor-based directional classifications and canonical rule-based heuristics commonly used in the hot-hand and gambler's-fallacy literature.

Analysis proceeds at two complementary levels. (i) At the *sequence level*, the core inferential models quantify how each sequence influences directional responding. (ii) At the *participant level*, descriptive summaries assess the economic consequences of these tendencies, including the extent to which potentially biased choices deviate from EU–optimal behavior. Together, these components provide a coherent structure for linking sequence-induced psychological patterns to their economic implications.

Sex will be treated as a descriptive characteristic of the sample. Sex-stratified descriptive summaries will be reported to characterize individual differences and aggregate behavioral patterns; sex is not included as a covariate in confirmatory models, and no sex-by-sequence interactions are estimated. For sequences identified as focal according to the preregistered criteria, additional sex-stratified descriptive summaries will be reported; sequence selection is based on overall sequence-level effects, not on observed sex differences.

All analyses are designed to be implementable end-to-end under a fully specified data-generating process. Prior to analyzing the experimental data, we will verify the internal coherence of the full estimation pipeline using simulated data generated from the assumed structural primitives (risk preferences, expected-utility–optimal stakes, and sequence-level choice tendencies). These simulations will assess numerical stability, identifiability of sequence- and participant-level effects, and the behavior of derived quantities under known ground truth. The purpose of these checks is not model selection or tuning, but to ensure that the preregistered estimands are well defined, computationally stable, and interpretable under plausible data realizations.

2.1 Confirmatory Analyses

This section specifies the preregistered confirmatory analyses conducted in the False–Negative treatment, where expected-utility–optimal behavior is well defined across all sequences. Each research question (RQ1–RQ4) targets a distinct behavioral margin and defines sequence-level estimands, model structure, and evidence criteria in advance. All confirmatory inference relies on within-participant variation and hierarchical pooling across sequences and participants; no model specifications or decision rules are introduced post hoc.

2.1.1 RQ1: Do any sequences systematically affect the likelihood of betting relative to the expected-utility–optimal choice?

This analysis isolates the extensive margin of decision-making—whether participants choose to bet or not after observing each sequence. It tests whether specific sequences systematically reduce the probability of betting relative to the expected-utility–optimal benchmark.

Bayesian estimands and descriptive posterior summaries. Under the rational-choice benchmark, betting is expected-utility–optimal in the FN treatment ($m = 2.5$) for all but extremely risk-averse participants. Systematic decreases in betting probability for some sequences therefore indicate deviations from expected-utility–consistent behavior. We summarise posterior mass below the benchmark using $U_s^b(\cdot) = P(\mu_s^b < 1 - \rho)$, with $\rho = 0.10$ (and sensitivity analyses for $\rho \in \{0.08, 0.12\}$). These probabilities provide descriptive measures of posterior concentration below the benchmark and do not constitute formal decision thresholds. Because the normative benchmark implies near-universal betting, even though observed behavior may deviate from this benchmark, even moderate departures from always betting (e.g., $\mu_s^b \approx 0.85$) are interpreted as economically meaningful under-betting rather than as negligible noise. Confirmatory analyses in RQ1 are conducted on the subset of participants for whom the optimal stake in the FN treatment is strictly positive with high posterior probability; the formal definition of this benchmark-consistent subset is provided in Appendix A.6.

Model specification. Let $\mathtt{b}_{is} \in \{0,1\}$ denote whether participant i placed a bet in sequence s (1 = bet placed, 0 = no bet). A Bayesian hierarchical logistic regression with a Bernoulli likelihood is fitted to estimate the probability of betting:

$$\mathtt{b}_{is} \sim \text{Bernoulli}(\pi^{b}_{is}), \qquad \text{logit}(\pi^{b}_{is}) = \alpha + u_i + \beta_s, \qquad u_i \sim \mathcal{N}(0, \sigma_u), \qquad \beta_s \sim \mathcal{N}(0, \sigma_{\text{s}}),$$

with weakly informative priors

$$\alpha \sim \mathcal{N}(0, 1.5), \qquad \sigma_u, \sigma_{\text{s}} \sim \text{HalfNormal}(0, 1).$$

Inference focuses on the sequence-level mean betting probability

$$\mu^b_s = E_i[\pi^b_{is}],$$

where the expectation is taken across participants. Sequence effects $\{\beta_s\}$ are identified relative to the grand mean by imposing the constraint $\sum_s \beta_s = 0$, so that α represents the population-average log-odds of betting.

Model diagnostics and robustness. Posterior predictive checks will assess the adequacy of the Bernoulli likelihood for $\mathtt{b}_{is}$. If systematic lack of fit is detected (e.g., overdispersion in sequence–wise betting frequencies), a Beta–Binomial specification on aggregated sequence-level counts with identical priors and hierarchical structure will be estimated as a robustness check.

Sequence-level summaries. A practical equivalence region is centred at the rational benchmark $\mu^b_s = 1$. For interpretive clarity, sequences are described according to the magnitude of posterior concentration below this benchmark using $U^b_s(\cdot)$. For exposition, "strong", "moderate", "weak", and "neutral" under-betting correspond, respectively, to $U^b_s(\cdot)$ lying in [0.95,1], [0.80,0.95), [0.50,0.80), and below 0.50. These labels provide descriptive summaries of posterior concentration and do not function as inferential cutoffs; inference relies on the full posterior distributions. Posterior medians, 95% credible intervals, the number of trials, and sensitivity to ρ are reported. Finally, if overall betting rates are substantially below the normative benchmark, the quantities $U^b_s(\cdot)$ may approach one for most sequences; in such cases, interpretation will rely primarily on the posterior medians and credible intervals of μ^b_s rather than on the descriptive labels attached to $U^b_s(\cdot)$.

Participant-level summaries. For each participant, posterior means of random effects yield the average betting probability across sequences:

$$\mu^b_i = E_s\big[E(\pi^b_{is} \mid \text{data})\big].$$

Individual tendencies are summarised using posterior medians, 95% credible intervals, and $U^b_i(\cdot) = P(\mu^b_i < 1 - \rho)$, with $\rho = 0.10$. Values of $U^b_i(\cdot)$ close to one indicate substantial posterior concentration on under-betting; intermediate values reflect milder departures; and small values indicate little posterior support for deviation. For exposition, the descriptive terms "solid", "likely", "leaning", and "neutral" under-better correspond, respectively, to $U^b_i(\cdot)$ in [0.95,1], [0.90,0.95), [0.75,0.90), and below 0.75. These labels are used purely to aid interpretation and do not serve as inferential criteria. Full posterior summaries of μ^b_i are reported.

2.1.2 RQ2: Do any sequences lead to over- or under-staking relative to the expected-utility–optimal amount?

This analysis focuses on the intensive margin of behavior—how much participants stake when they choose to bet. It examines whether specific sequences systematically shift stake calibration relative to the expected-utility–optimal amount, thereby revealing distortions in the scale of risk taking rather than in the decision to participate.

Bayesian estimands and descriptive posterior summaries. Under the rational-choice benchmark, mean stake deviations should be close to zero. Posterior mass outside a practical equivalence region is summarised using $U_s^a(\cdot) = P(\mu_s^a < -\rho)$ for under-staking and $O_s^a(\cdot) = P(\mu_s^a > \rho)$ for over-staking, with $\rho = 0.05$ (and sensitivity analyses using $\rho \in \{0.03, 0.08\}$). These probabilities describe posterior concentration on calibration errors and do not constitute formal decision thresholds. Confirmatory analyses in RQ2 are estimated on the subset of participants with sufficient observed betting variation and for whom betting is optimal in the FN treatment (Appendix A.6).

Model specification. The expected-utility–optimal amount a_{is}^* is defined in the *Parameter definitions* section and computed for each posterior draw of r_i. Stake deviations are defined as

$$\Delta a_{is} = a_{is} - a_{is}^*.$$

To isolate sequence-level effects from individual differences in staking scale, deviations are normalised within participant:

$$Z_{is}^{\Delta a} = \frac{\Delta a_{is} - \overline{\Delta a}_i}{s_i^*(\Delta a)}, \qquad s_i^*(\Delta a) = \max(\mathrm{sd}_i(\Delta a),\, 2),$$

where $\overline{\Delta a}_i$ and $\mathrm{sd}_i(\Delta a)$ are computed across betting trials of participant i. The regularised denominator stabilises the transformation when participants stake nearly constant amounts. Participants with fewer than three betting trials provide insufficient within-participant variation for the construction of $Z_{is}^{\Delta a}$ and are therefore excluded from the RQ2 analysis. Their exclusion will be noted in the reporting tables.

A Bayesian hierarchical model is fitted:

$$Z_{is}^{\Delta a} = \alpha + u_i + \beta_s + \varepsilon_{is}, \qquad u_i \sim \mathcal{N}(0, \sigma_u), \quad \beta_s \sim \mathcal{N}(0, \sigma_{\mathrm{s}}), \quad \varepsilon_{is} \sim \mathcal{N}(0, \sigma),$$

with weakly informative priors

$$\alpha \sim \mathcal{N}(0, 1), \qquad \sigma_u,\, \sigma_{\mathrm{s}},\, \sigma \sim \mathrm{HalfNormal}(0, 1).$$

For interpretive convenience, posterior sequence effects are mapped back to an absolute scale to obtain a descriptive sequence-level calibration contrast:

$$\mu_s^a = \frac{1}{e}\, E_i\big[\overline{\Delta a}_i + (\alpha + u_i + \beta_s)\, s_i^*(\Delta a)\big],$$

so that results remain interpretable as percentage deviations from the expected-utility–optimal amount. Because this mapping combines participant-level average deviations with sequence-level effects rescaled by participant-specific dispersion measures, μ_s^a should accordingly be interpreted as a descriptive contrast rather than as a structural population mean; its absolute magnitude depends on the chosen regularization of within-participant dispersion.

Model diagnostics and robustness. If posterior predictive checks detect boundary inflation caused by all-in choices ($a = 100$), a one-inflated Beta–Binomial model on the 1–100 scale will be estimated using the same hierarchical structure and priors. As robustness checks, the analysis will be repeated using alternative floors $s_i^*(\Delta a) = \max(\mathrm{sd}_i(\Delta a), c)$ with $c \in \{1, 5\}$, and using a robust dispersion measure $s_i = \mathrm{MAD}_i(\Delta a)$ (median absolute deviation) with the same floor.

Sequence-level summaries. For interpretive clarity, descriptive labels indicate the magnitude and direction of posterior concentration: for under-staking, $U_s(\cdot)$ in $[0.95, 1]$ (strong), $[0.80, 0.95)$ (moderate), $[0.50, 0.80)$ (weak), and below 0.50 (neutral); for over-staking, the same intervals apply to $O_s(\cdot)$. These labels aid exposition but do not serve as inferential criteria. Posterior medians, 95% credible intervals, the number of betting trials, and sensitivity to ρ are reported.

Participant-level summaries. Average calibration per participant is defined as

$$\mu_i^a = E_s[E(\Delta a_{is}/e \mid \text{data})].$$

Posterior summaries (medians, 95% credible intervals, and $U_i^a(\cdot) = P(\mu_i^a < -\rho)$ and $O_i^a(\cdot) = P(\mu_i^a > \rho)$, with $\rho = 0.05$) describe individual tendencies toward under- or over-staking. Values of $U_i^a(\cdot)$ or $O_i^a(\cdot)$ close to one indicate substantial posterior concentration on under- or over-staking; intermediate values reflect milder departures; and small values indicate little posterior support for deviation. The descriptive labels "solid", "likely", "leaning", and "neutral" correspond to the following intervals: for under-staking, $U_i^a(\cdot)$ in $[0.95, 1]$, $[0.90, 0.95)$, $[0.75, 0.90)$, and below 0.75; for over-staking, the same intervals apply to $O_i^a(\cdot)$. These labels aid interpretation and do not function as inferential criteria. Full posterior summaries of μ_i^a are reported.

2.1.3 RQ3: Do any sequences increase certainty-equivalent losses from suboptimal betting?

Building on RQ1 and RQ2, this analysis integrates extensive-margin errors (whether to bet) and intensive-margin deviations (how much to bet) into a unified welfare metric. It quantifies the total economic cost of suboptimal decisions in expected-utility terms, capturing the combined effects of false positives, false negatives, and stake miscalibration as monetarily interpretable certainty-equivalent losses.

Bayesian estimands and descriptive posterior summaries. Under the rational-choice benchmark, welfare losses should be negligible. Systematic increases in certainty-equivalent loss indicate departures from expected-utility–optimal decision-making. Posterior mass above a practical equivalence band is summarised using $L_s(\cdot) = P(\mu_s^c > \rho)$, with $\rho = 0.05$ (and sensitivity analyses for $\rho \in \{0.03, 0.08\}$). These probabilities describe posterior concentration above the benchmark and do not constitute formal decision thresholds. Confirmatory welfare-loss analyses are restricted to participants for whom betting is optimal in the FN treatment; see Appendix A.6 for the formal definition.

Model specification. The certainty equivalent (CE) is defined in Section A.6. Welfare loss from suboptimal betting is computed as

$$\Delta c_{is} = CE(a_{is}^*; r_i) - CE(a_{is}; r_i),$$

where a_{is}^* is the expected-utility–optimal stake for participant i in sequence s. Positive values indicate welfare losses relative to the optimal decision; small negative values may arise from numerical noise when a_{is} is adjacent to a_{is}^* and posterior draws of r_i vary. For modeling purposes, all values $\Delta c_{is} \leq 0$ are deterministically set to zero prior to estimation and treated as zero welfare loss.

Because welfare loss is non-negative by construction and equals zero whenever the chosen stake coincides with the expected-utility–optimal stake, while positive losses are continuous, bounded above by a finite constant determined by the utility function and endowment, and may be right-skewed, the primary likelihood is a *hurdle–Gamma model.* The hurdle component captures the probability of zero loss if such exact matches occur, and the Gamma component models the distribution of strictly positive losses:

$$\Delta c_{is} \sim \begin{cases} 0, & \text{with probability } \pi_{is}^0, \\ \text{Gamma}(k_{is}, \theta_{is}), & \text{with probability } 1 - \pi_{is}^0, \end{cases}$$

with

$$\text{logit}(\pi_{is}^0) = \alpha^{(0)} + u_i^{(0)} + \beta_s^{(0)}, \qquad \log \mathbb{E}[\Delta c_{is} \mid \Delta c_{is} > 0] = \alpha^{(+)} + u_i^{(+)} + \beta_s^{(+)}.$$

All random effects have Normal priors with weakly informative HalfNormal priors on their scales. The data determine whether the hurdle probability is negligible or non-negligible.

Sequence-level welfare loss is summarised as

$$\mu_s^c = E_i[\Delta c_{is}/e],$$

expressed as a share of the endowment. A practical equivalence region is centred at zero with threshold $\rho = 0.05$.

Model diagnostics and robustness. As a simplified working alternative, a Gaussian hierarchical model with the same random-effects structure is also fitted:

$$\Delta c_{is} = \alpha + u_i + \beta_s + \varepsilon_{is}, \qquad \varepsilon_{is} \sim \mathcal{N}(0, \sigma),$$

with weakly informative priors $\alpha \sim \mathcal{N}(0, 1)$ and σ, σ_u, $\sigma_\mathrm{s} \sim \mathrm{HalfNormal}(0, 1)$. Posterior predictive checks compare the Gaussian and hurdle–Gamma specifications. If the estimated hurdle probability is negligible and posterior predictive distributions show no substantial point mass at zero, results will additionally be reported using a simplified continuous specification without a hurdle (Gamma or log-normal likelihood with the same hierarchical structure). The Gaussian model does not naturally accommodate the boundary at zero or potential right skew but provides a useful diagnostic for the sign and relative ordering of sequence effects. Substantive conclusions rely on the stability of posterior summaries across likelihood specifications rather than on any single distributional assumption.

Sequence-level summaries. For interpretive clarity, sequences are described according to the magnitude of posterior concentration above the practical equivalence region using $L_s(\cdot)$. For exposition, the descriptive labels "strong", "moderate", "weak", and "neutral" correspond respectively to $L_s(\cdot)$ in $[0.95, 1]$, $[0.80, 0.95)$, $[0.50, 0.80)$, and below 0.50. These labels provide qualitative summaries and do not serve as inferential criteria. Posterior medians, 95% credible intervals, the number of trials, and sensitivity to ρ are reported.

Participant-level summaries. Average welfare loss per participant is defined as

$$\mu_i^c = E_s[E(\Delta c_{is}/e \mid \text{data})].$$

Posterior summaries (medians, 95% credible intervals, and $L_i(\cdot) = P(\mu_i^c > \rho)$ with $\rho = 0.05$) describe individual decision efficiency. Values of $L_i(\cdot)$ close to one indicate substantial posterior concentration on welfare loss; intermediate values reflect milder departures; and small values indicate little posterior support for inefficiency. The descriptive labels "solid", "likely", "leaning", and "neutral" correspond to $L_i(\cdot)$ in $[0.95, 1]$, $[0.90, 0.95)$, $[0.75, 0.90)$, and below 0.75, respectively. These labels aid interpretation and do not serve as inferential categories. Full posterior summaries of μ_i^c are reported.

2.1.4 RQ4: Do any sequences systematically bias side choices toward Heads or Tails?

This analysis examines whether any sequence systematically shifts side-choice probabilities conditional on betting, away from the participant-level baseline tendency to choose Heads or Tails.

Bayesian estimands and descriptive posterior summaries. Under the rational-choice benchmark, there is no normative reason to prefer Heads or Tails. Any individual may exhibit a stable personal tendency, but at the population level the baseline probability of choosing Heads is defined as the population-average choice probability implied by the intercept and participant heterogeneity:

$$\bar{h} = E_{u \sim \mathcal{N}(0, \sigma_u)} [\mathrm{logistic}(\alpha + u)] .$$

This baseline is expected to remain stable across sequences. Sequence-specific means μ_s^h are therefore expected to cluster around $\bar{h}$. Posterior mass outside a practical tolerance band is summarised using $H_s(\cdot) = P(\mu_s^h > \bar{h} + \delta)$ for Head-bias and $T_s(\cdot) = P(\mu_s^h < \bar{h} - \delta)$ for Tail-bias, with $\delta = 0.05$ and sensitivity analyses using $\delta \in \{0.03, 0.08\}$. These probabilities describe posterior concentration on Head- or Tail-biased responding and do not constitute formal decision thresholds.

Model specification. Side-choice analyses in RQ4 and the subsequent exploratory analysis EX1 are conducted only on trials in which a bet is placed ($b_{is} = 1$) and are therefore interpreted as conditional choice probabilities $P(h_{is} \mid b_{is} = 1)$; no unconditional claims about side preferences are made. Let $\mathtt{h}_{is} \in \{0, 1\}$ denote whether participant i chose Heads (1) or Tails (0). A Bayesian hierarchical logistic regression with a Bernoulli likelihood models the probability of choosing Heads:

$$\mathtt{h}_{is} \sim \text{Bernoulli}(\pi_{is}^h), \qquad \text{logit}(\pi_{is}^h) = \alpha + u_i + \beta_s, \qquad u_i \sim \mathcal{N}(0, \sigma_u), \qquad \beta_s \sim \mathcal{N}(0, \sigma_\text{s}),$$

with $\pi_{is}^h = P(\mathtt{h}_{is} = 1)$ and weakly informative priors

$$\alpha \sim \mathcal{N}(0, 1.5), \qquad \sigma_u, \sigma_\text{s} \sim \text{HalfNormal}(0, 1).$$

The sequence-level mean choice probability is

$$\mu_s^h = E_i[\pi_{is}^h],$$

and deviations of μ_s^h from $\bar{h}$ identify sequence-induced directional bias. The sequence effects $\{\beta_s\}$ are identified relative to the grand mean by imposing the constraint $\sum_s \beta_s = 0$, so that α represents the population-average log-odds of choosing Heads.

Model diagnostics and robustness. Posterior predictive checks will assess the adequacy of the Bernoulli likelihood for $\mathtt{h}_{is}$. If posterior predictive distributions exhibit overdispersion in sequence-wise Head-choice frequencies, a Beta–Binomial specification on aggregated sequence-level counts, with the same hierarchical structure and priors, will be estimated as a robustness check.

Sequence-level summaries. Sequence-level directional biases are described on a seven-point directional scale ranging from strong Tail-bias to strong Head-bias. Given $H_s(\cdot) = P(\mu_s^h > \bar{h} + \delta)$ and $T_s(\cdot) = P(\mu_s^h < \bar{h} - \delta)$, sequences are labelled "strong", "moderate", or "weak" Head-biased when $H_s(\cdot)$ lies in $[0.95, 1]$, $[0.80, 0.95)$, or $[0.50, 0.80)$, respectively, and analogously labelled "weak", "moderate", or "strong" Tail-biased when $T_s(\cdot)$ falls in the same intervals. Sequences for which both $H_s(\cdot)$ and $T_s(\cdot)$ are below 0.50 are described as "neutral". These terms are purely descriptive; all inference relies on the full posterior distributions. Posterior medians, 95% credible intervals, the number of trials, and sensitivity to δ are reported. For transparency, $P(|\mu_s^h - 0.5| > \delta)$ is additionally reported, capturing absolute deviations from parity.

Participant-level summaries. Participant-level tendencies are defined as

$$\mu_i^h = E_s[E(\pi_{is}^h \mid \text{data})], \qquad \bar{h} = \text{logistic}(\alpha).$$

Posterior summaries of μ_i^h (medians, 95% credible intervals, $H_i(\cdot) = P(\mu_i^h > \bar{h} + \delta)$ and $T_i(\cdot) = P(\mu_i^h < \bar{h} - \delta)$) are used to describe individual directional biases. Participants are labelled "solid", "likely", or "leaning" "Headish" when $H_i(\cdot)$ lies in $[0.95, 1]$, $[0.90, 0.95)$, or $[0.75, 0.90)$, respectively, and analogously labelled "solid", "likely", or "leaning" "Tailish" when $T_i(\cdot)$ falls in the same intervals. Participants with both $H_i(\cdot)$ and $T_i(\cdot)$ below 0.75 are described as "neutral". As with sequences, these labels are descriptive groupings; interpretation is based on the full posterior distributions. For transparency, $P(|\mu_i^h - 0.5| > \delta)$ is also reported.

2.2 Exploratory Analyses

Exploratory analyses extend the confirmatory framework by constructing derived indices, examining associations with individual characteristics, assessing robustness across incentive environments, and comparing anchor-based classifications to canonical heuristic measures. These analyses are descriptive and hypothesis-generating; they do not alter confirmatory estimands or evidence criteria defined in Section 2.1.

2.2.1 EX1: Hot-handish and gamblerish patterns in side choice

EX1 extends the side-choice findings from RQ4 by examining whether certain sequences systematically generate hot-handish (streak-following) or gamblerish (streak-opposing) choices. Traditional research on hot-hand and gambler's fallacies typically relies on ex–ante assumptions about what constitutes a streak (e.g., focusing only on the last outcome, counting alternations, or using the length of the final run). To avoid such assumptions, EX1 defines streaks empirically through distributional similarity to two behavioral anchors.

The pure sequences `HHHHHH` and `TTTTTT` are the only cases in which choosing the same side is unambiguously hot-handish and choosing the opposite side unambiguously gamblerish. Posterior Heads/Tails choice distributions of all impure sequences are compared to those of the two anchors and to a neutral baseline. This comparison yields a set of weights that probabilistically determine whether choosing Heads or Tails on any given sequence reflects hot-handish, gamblerish, or neutral responding. Averaging these directional scores across participants produces sequence-level directional indices, which allow us to identify sequences that behave similarly to the pure anchors or remain neutral.

Averaging directional scores across sequences instead yields participant-level directional tendencies. These define the Gamblerish–Hot-handish Index (GHI; ranging from -1 to 1), where positive values indicate hot-handishness and negative values indicate gamblerishness. The second part of EX1 tests whether optimism, response times, or risk preferences predict variation in the GHI. The third part evaluates whether a simple score based solely on responses to the pure sequences approximates the GHI derived from the full sequence set.

I. Distributional Similarity and the Gamblerish–Hot-handish Index

For each sequence s, the hierarchical RQ4 model yields posterior draws of the population Heads-choice probability

$$\theta_s^{(k)} = \mu_s^{h\,(k)} = E_i[\pi_{is}^{h\,(k)}], \qquad k = 1, \ldots, K.$$

These draws represent posterior uncertainty about the population tendency to choose Heads when betting on sequence s. Here K denotes the number of retained posterior draws.

Two sequences serve as behavioral anchors,

$$\bar{H} = \texttt{HHHHHH}, \qquad \bar{T} = \texttt{TTTTTT},$$

with posterior draws $\{\theta_{\bar{H}}^{(k)}\}$ and $\{\theta_{\bar{T}}^{(k)}\}$ obtained from the RQ4 model. On $\bar{H}$, choosing Heads is interpreted as hot-handish and choosing Tails as gamblerish; on $\bar{T}$, the directional meaning is reversed. A neutral baseline reference θ_0 is obtained from the RQ4 model and is defined, for each posterior draw k, as the population-average probability of choosing Heads in the absence of sequence-specific effects:

$$\theta_0^{(k)} = E_{u \sim \mathcal{N}(0, \sigma_u^{(k)})}\left[\text{logistic}\big(\alpha^{(k)} + u\big)\right],$$

where $\alpha^{(k)}$ and $\sigma_u^{(k)}$ are the intercept and participant-level standard deviation from the RQ4 model.

Posterior similarity to behavioral anchors. To quantify posterior similarity, posterior draws of the sequence-level choice probabilities are compared directly. For each sequence s and anchor $a \in \{\bar{H}, \bar{T}, 0\}$, similarity is defined as the posterior probability that the population Heads-choice probability for s is close to that of the anchor:

$$d_a(s) = P(|\theta_s - \theta_a| < \epsilon \,|\, \text{data}) \,,$$

where $\epsilon > 0$ is a tolerance parameter. These probabilities are estimated by pairing posterior draws across MCMC iterations,

$$d_a(s) \approx \frac{1}{K} \sum_{k=1}^{K} \mathbf{1}\Big\{\Big|\theta_s^{(k)} - \theta_a^{(k)}\Big| < \epsilon\Big\} \,.$$

The quantities $d_a(s)$ lie in $[0, 1]$ and represent posterior similarity.

Anchor-based similarity weights. Similarity weights are obtained by normalising the posterior similarity probabilities:

$$w_a(s) = \frac{d_a(s) + \eta}{d_{\bar{H}}(s) + d_{\bar{T}}(s) + d_0(s) + 3\eta}, \qquad a \in \{\bar{H}, \bar{T}, 0\},$$

with a small stabilizing constant $\eta = 10^{-6}$. By construction,

$$w_{\bar{H}}(s) + w_{\bar{T}}(s) + w_0(s) = 1.$$

The main analysis uses $\epsilon = 0.05$, with sensitivity analyses for $\epsilon \in \{0.03, 0.08\}$.

These weights determine how directional meaning is assigned on each sequence.

Directional meaning of individual choices. For each trial (i, s) with observed side choice $h_{is} \in \{0, 1\}$ ($1 =$ Heads, $0 =$ Tails),

$$P(\text{hot} \mid h_{is}, s) = \begin{cases} w_{\bar{H}}(s) \cdot 1 + w_{\bar{T}}(s) \cdot 0 + w_0(s) \cdot 0.5, & h_{is} = 1, \\ w_{\bar{H}}(s) \cdot 0 + w_{\bar{T}}(s) \cdot 1 + w_0(s) \cdot 0.5, & h_{is} = 0, \end{cases}$$

and $P(\text{gambler} \mid h_{is}, s) = 1 - P(\text{hot} \mid h_{is}, s)$.

A continuous trial-level directional score is defined as

$$z_{is} = P(\text{hot} \mid h_{is}, s) - P(\text{gambler} \mid h_{is}, s), \qquad z_{is} \in [-1, 1].$$

Values near $+1$ indicate strongly hot-handish responding; values near -1 indicate strongly gamblerish responding; values near 0 indicate ambiguous or neutral responding.

Sequence-level directional tendencies. Sequence-level directional tendencies are defined as

$$\chi_s = E_i[z_{is}],$$

with posterior draws obtained from the joint posterior distribution of the $\{z_{is}\}$.

Sequence-level summaries. Sequence-level directional similarity is classified on a seven-point scale ranging from strong $\bar{T}$-similarity to strong $\bar{H}$-similarity. Let $H(\cdot) = P(\chi_s > \delta)$ and $T(\cdot) = P(\chi_s < -\delta)$, with $\delta = 0.05$ (and sensitivity analyses at $\delta \in \{0.03, 0.08\}$). Sequences are labeled *strong*, *moderate*, or *weak* $\bar{H}$-similar when $H(\cdot)$ lies in [0.95,1], [0.80,0.95), or [0.50,0.80), respectively, and analogously labeled *weak*, *moderate*, or *strong* $\bar{T}$-similar when $T(\cdot)$ lies in these intervals. Sequences for which both probabilities fall below 0.50 are classified as *neutral*. Posterior medians and 95% credible intervals for χ_s, as well as $H(\cdot)$, $T(\cdot)$, and $P(|\chi_s| > \delta)$, are reported.

Participant-level directional tendencies. Participant-level directional tendencies are defined as the across-sequence average of directional scores:

$$\chi_i = E_s[z_{is}],$$

the participant-level *Gamblerish–Hot-handish Index* (GHI). Positive values of χ_i indicate hot-handish responding and negative values indicate gamblerish responding.

Posterior probabilities of directional tendency are

$$HH(\cdot) = P(\chi_i > \delta) \qquad \text{and} \qquad G(\cdot) = P(\chi_i < -\delta),$$

with the same equivalence threshold $|\chi_i| \leq \delta$. Participants are classified as *solid*, *likely*, or *leaning* hot-handish when $HH(\cdot)$ lies in [0.95,1], [0.90,0.95), or [0.75,0.90), respectively, and analogously as *leaning*, *likely*, or *solid* gamblerish when $G(\cdot)$ lies in [0.75,0.90), [0.90,0.95), or [0.95,1]. Participants for whom both probabilities fall below 0.75 are classified as *neutral*. Posterior medians, 95% credible intervals, and the probabilities $HH(\cdot)$, $G(\cdot)$, and $P(|\chi_i| > \delta)$ are reported.

II. Predictors of the GHI

This section examines whether individual characteristics predict directional tendencies. The dependent variable is the participant-level Gamblerish–Hot-handish Index (GHI). Because the GHI is a derived posterior quantity from Section 2.2.1, all regression analyses propagate this uncertainty.

For each posterior draw $t = 1, \ldots, T$, we compute a draw of participant-level directional tendency as the across-sequence average of trial-level directional scores:

$$\chi_i^{(t)} = E_s[z_{is}^{(t)}].$$

For each draw t, we then estimate the following Bayesian regression using $\{\chi_i^{(t)}\}_{i=1}^N$ as the dependent variable:

$$\chi_i^{(t)} = \alpha + \beta_{\text{opt}}\, Z_i^{\text{opt}} + \beta_{\text{rt}}\, Z_i^{\text{rt}} + \beta_r\, Z_i^r + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma_\varepsilon).$$

Predictors are standardized across participants: Z_i^{opt} is standardized LOT–R, Z_i^{rt} is standardized log response time, and Z_i^r is standardized risk parameter.

Weakly informative priors are used:

$$\alpha,\, \beta_{\text{opt}},\, \beta_{\text{rt}},\, \beta_r \sim \mathcal{N}(0, 1), \qquad \sigma_\varepsilon \sim \text{HalfNormal}(0, 1).$$

Posterior summaries of regression coefficients are obtained by aggregating information across draws t. We report posterior medians, 95% credible intervals, and posterior probabilities of direction $P(\beta > 0)$ for all coefficients.

III. Pure-Sequence Approximation to the GHI

Because the full GHI uses all 64 sequences and incorporates uncertainty via distributional similarity, it is useful to assess whether a simpler measure based solely on the pure sequences approximates the same directional tendencies.

For each participant i, define the pure-sequence analogue of the Gamblerish–Hot-handish Index:

$$\chi_i^{\text{pure}} = \mathbf{1}\{h_{i,\bar{H}} = 1\} - \mathbf{1}\{h_{i,\bar{T}} = 1\}, \qquad \chi_i^{\text{pure}} \in \{-1,\, 0,\, 1\}.$$

Here, $\chi_i^{\text{pure}} = 1$ indicates hot-handish responses on both pure sequences, $\chi_i^{\text{pure}} = -1$ indicates gamblerish responses on both, and $\chi_i^{\text{pure}} = 0$ indicates mixed responding.

We summarize the association between χ_i^{pure} and the full Gamblerish–Hot-handish Index via the posterior distribution of the correlation

$$\rho_{\text{pure},\chi} = \text{cor}(\chi_i^{\text{pure}},\, \chi_i),$$

reporting posterior medians, 95% credible intervals, and $P(\rho_{\text{pure},\chi} > 0)$. Posterior distributions of χ_i are also reported separately for the three pure-sequence categories $\chi_i^{\text{pure}} \in \{-1, 0, 1\}$.

This comparison assesses whether the two pure sequences provide a sufficient behavioral summary of directional responding or whether the distributional information from the full sequence space adds meaningful predictive structure.

2.2.2 EX2: Participant-level associations with optimism and response time across RQ1–RQ4

This exploratory analysis examines whether dispositional optimism (LOT–R) and mean response time predict participant-level behavioral patterns across the confirmatory outcomes RQ1–RQ4. The standardized predictors Z_i^{opt} (optimism) and Z_i^{rt} (log-transformed response time) are defined as in EX1.

Participant-level outcomes are taken from the posterior estimates of the confirmatory models in the False–Negative treatment:

$$\left\{\mu_i^b,\ \mu_i^a,\ \mu_i^c,\ \mu_i^h\right\},$$

where μ_i^b is the individual betting probability from RQ1, μ_i^a is the mean proportional stake deviation from RQ2, μ_i^c is the mean certainty-equivalent loss from RQ3, and μ_i^h is the baseline probability of choosing Heads from RQ4.

Some participant-level outcomes are undefined for subsets of participants due to limited observed betting. In particular, μ_i^a is undefined for participants with fewer than three betting trials in the FN treatment. Analyses involving a given outcome are therefore conducted on the subset of participants for whom that outcome is well-defined; participants are not excluded globally and may contribute to other outcome regressions within EX2.

Each quantity enters the regression via posterior draws and is standardized within draw across participants to ensure comparability across outcomes. Let y_{ki} denote the standardized posterior draw for outcome $k \in \{b, a, c, h\}$ of participant i. A multivariate Bayesian regression with partial pooling across outcomes is estimated:

$$y_{ki} = \alpha_k + \beta_{k,\text{opt}}\, Z_i^{\text{opt}} + \beta_{k,\text{rt}}\, Z_i^{\text{rt}} + \varepsilon_{ki}, \qquad \varepsilon_{ki} \sim \mathcal{N}(0, \sigma_k).$$

To borrow strength across outcomes, slopes are partially pooled:

$$\beta_{k,\text{opt}} \sim \mathcal{N}(\bar{\beta}_{\text{opt}},\ \tau_{\text{opt}}), \qquad \beta_{k,\text{rt}} \sim \mathcal{N}(\bar{\beta}_{\text{rt}},\ \tau_{\text{rt}}).$$

Weakly informative priors are used:

$$\bar{\beta}_{\mathrm{opt}},\ \bar{\beta}_{\mathrm{rt}} \sim \mathcal{N}(0,1), \qquad \tau_{\mathrm{opt}},\ \tau_{\mathrm{rt}} \sim \mathrm{HalfNormal}(0,0.5),$$

$$\alpha_k \sim \mathcal{N}(0,1), \qquad \sigma_k \sim \mathrm{HalfNormal}(0,1).$$

A robustness specification replaces independent residuals with multivariate Normal residuals and an LKJ(2) prior on the residual correlation matrix.

Posterior medians, 95% credible intervals, and posterior probabilities of direction $P(\cdot > 0)$ will be reported for the pooled slopes $(\bar{\beta}_{\mathrm{opt}}, \bar{\beta}_{\mathrm{rt}})$ and the outcome-specific slopes $(\beta_{k,\mathrm{opt}}, \beta_{k,\mathrm{rt}})$. Model estimation, convergence diagnostics, and posterior predictive checks follow the general procedures described in Appendix A.

2.2.3 EX3: False–Positive treatment replication and FN–FP contrasts

The False–Positive (FP; $m = 1.9$) treatment provides an incentive frame in which betting is generally expected to be avoided, offering a complementary environment to the FN treatment used for confirmatory inference. Although the two treatments differ in their normative structure, examining behavior in FP allows us to assess the stability of sequence-level and participant-level patterns across markedly different incentives, identify which components of decision-making generalize across frames, and evaluate whether the patterns observed in FN reflect context-dependent responses or broader behavioral tendencies.

This section follows the structure of RQ1–RQ4 and examines, for each component, which aspects can be meaningfully replicated in FP and which can be contrasted across treatments. All analyses in EX3 are descriptive and do not introduce additional inferential criteria. All FP models use the same likelihoods, priors, drift adjustments, and diagnostic procedures specified in Section A.

RQ1 (Extensive margin of betting). The hierarchical betting model from RQ1 is re-estimated in the FP treatment using the same likelihood, priors, drift adjustments, and estimation settings as in FN. In FP, the model identifies over-betting (false positives), mirroring the under-betting (false negatives) examined in FN.

For each sequence s, let $\mu_s^{b,\mathrm{FP}}$ denote the posterior mean betting probability in FP. FN–FP contrasts are computed as

$$\Delta_s^b = \mu_s^{b,\mathrm{FN}} - \mu_s^{b,\mathrm{FP}},$$

with posterior medians, 95% credible intervals, and $\Pr(\Delta_s^b > 0)$ reported. Descriptive thresholds are also summarized via $\Pr(|\Delta_s^b| > 0.05)$. At the participant level, betting differences are summarized as

$$\Delta_i^b = \mu_i^{b,\mathrm{FN}} - \mu_i^{b,\mathrm{FP}},$$

with posterior medians, 95% credible intervals, and $\Pr(\Delta_i^b > 0)$ reported.

RQ2 (Intensive margin of staking). The intensive margin from RQ2 is expected to be non-estimable at the sequence level in FP: the expected-utility–optimal stake is $a^* = 0$ for nearly all participants, so FP will likely yield too few positive stakes for hierarchical calibration. Nevertheless, any positive FP stakes offer diagnostic information on subjective win beliefs.

For each FP trial with $a_{is} > 0$, the implied subjective probability that rationalizes the observed stake is computed by solving, for each posterior draw of r_i,

$$\frac{\partial EU(a_{is};\, p, r_i)}{\partial a} = 0,$$

for $p \in [0, 1]$, yielding $\hat{p}_{is}$.

If the first-order condition admits no interior solution on $[0, 1]$—including cases arising from structural corner solutions under sufficiently risk-loving preferences— or if numerical root-finding fails to converge or returns a value outside $[0, 1]$, the implied probability is set deterministically to the closest boundary value $\hat{p}_{is} \in \{0, 1\}$. All such cases are flagged. Participant-level summaries are reported both including and excluding flagged boundary cases as a robustness check. Posterior medians, 95% credible intervals, and $\Pr(\hat{p}_{is} > 0.5)$ are reported as participant-level sensitivity measures.

RQ3 (Welfare consequences of suboptimal betting). The welfare-loss model from RQ3 is re-estimated in the FP treatment using the same hierarchical structure and priors as in FN:

$$\Delta c_{is} = CE(a^*_{is}; r_i) - CE(a_{is}; r_i), \qquad \mu_s^{c,\mathrm{FP}} = E_i[\Delta c_{is}/e].$$

If posterior predictive checks indicate point mass near zero, the hurdle–Gamma robustness specification from RQ3 is applied.

Sequence-level FN–FP contrasts are computed as

$$\Delta_s^c = \mu_s^{c,\mathrm{FN}} - \mu_s^{c,\mathrm{FP}},$$

with posterior medians, 95% credible intervals, $\Pr(\Delta_s^c > 0)$, and $\Pr(|\Delta_s^c| > \rho_\Delta)$ for $\rho_\Delta \in \{0.03, 0.05\}$ reported. Spearman correlations between the FN and FP vectors $\{\mu_s^c\}$ are also reported.

RQ4 (Directional bias in side choice). Side-choice behavior is modeled in FP using the same hierarchical logistic structure as in FN, yielding posterior sequence-level choice probabilities $\mu_s^{h,\mathrm{FP}}$. However, FN–FP contrasts are not computed for RQ4: side choices in FP are observed only on trials where participants choose to bet, and betting probabilities differ sharply across incentive frames. This generates treatment-specific selection into the conditioning set $P(h_{is} \mid b_{is} = 1)$, making sequence-level and participant-level differences in side-choice frequencies across FN and FP behaviorally non-comparable. FP side-choice estimates are therefore reported descriptively (posterior medians, credible intervals, and deviation from the global FP baseline), but not contrasted with FN values.

2.2.4 EX4: Similarity Between Anchor-Based and Canonical Rule-Based Classifications

This exploratory analysis examines how closely canonical rule-based classifications commonly used in the hot-hand and gambler's-fallacy literature correspond to the anchor-based directional measures defined in EX1. Similarity is assessed at both the sequence level and the participant level. Because the anchor-based directional tendencies χ_s and χ_i are posterior-derived quantities (Section 2.2.1), all similarity statistics propagate posterior uncertainty.

Canonical rule-based classification and scoring. We consider a prespecified set of canonical heuristic rules: (i) the last-outcome rule; (ii) run-length–conditioned last-outcome rules (terminal run length $\ell_s \geq 2$ and $\ell_s \geq 3$); (iii) an imbalance-based rule reflecting law-of-small-numbers intuition; and (iv) an alternation-based rule. For each betting trial (i, s) and rule k, the observed side choice is classified as hot-handish ($+1$) or gamblerish (-1), yielding indicators

$$q_{is}^{(k)} \in \{-1, +1\}.$$

When a rule is undefined for a given sequence, the corresponding trial is omitted from aggregation by construction.

Formal definition of canonical rules. Let the six-toss sequence s be denoted by $(x_{s1}, \dots, x_{s6})$, where $x_{st} \in \{+1, -1\}$ corresponds to Heads and Tails, respectively. Let $c_{is} \in \{+1, -1\}$ denote participant i's chosen side on sequence s.

- *Last-outcome rule.*
$$q_{is}^{(\text{last})} = c_{is} \cdot x_{s6}.$$

- *Run-length–conditioned rule (threshold $\ell_s \geq 2$).* Let ℓ_s denote the length of the terminal run in sequence s.
$$q_{is}^{(\text{run2})} = \begin{cases} c_{is} \cdot x_{s6}, & \ell_s \geq 2, \\ \text{undefined}, & \ell_s < 2. \end{cases}$$

- *Run-length–conditioned rule (threshold $\ell_s \geq 3$).*
$$q_{is}^{(\text{run3})} = \begin{cases} c_{is} \cdot x_{s6}, & \ell_s \geq 3, \\ \text{undefined}, & \ell_s < 3. \end{cases}$$

- *Imbalance rule.* Let $B_s = \sum_{t=1}^{6} x_{st}$ denote sequence imbalance. Choices opposing the majority outcome are classified as gamblerish:
$$q_{is}^{(\text{imb})} = -\, c_{is} \cdot \operatorname{sign}(B_s),$$
with sequences satisfying $B_s = 0$ left unclassified.

- *Alternation rule.* Let $A_s = \sum_{t=2}^{6} \mathbf{1}\{x_{st} \neq x_{s,t-1}\}$ denote the number of alternations. Sequences with relatively few alternations are treated as streak-like:
$$q_{is}^{(\text{alt})} = \begin{cases} c_{is} \cdot x_{s6}, & A_s \leq 2, \\ -\, c_{is} \cdot x_{s6}, & A_s \geq 4, \\ \text{undefined}, & A_s = 3. \end{cases}$$

Sequence-level similarity. For each rule k, trial-level classifications are aggregated across participants to form a sequence-level rule-based score
$$Q_s^{(k)} = \frac{1}{N_s^{(k)}} \sum_i q_{is}^{(k)},$$
where $N_s^{(k)}$ denotes the number of included betting observations for sequence s under rule k.

Let $\chi_s^{(t)}$ denote posterior draws of the anchor-based sequence-level directional tendency from EX1, $t = 1, \dots, T$. For each rule k and posterior draw t, overall directional agreement is computed as
$$\text{Agree}_{\text{seq}}^{(k,t)} = \frac{1}{64} \sum_s \mathbf{1}\Big\{\operatorname{sign}\big(Q_s^{(k)}\big) = \operatorname{sign}\big(\chi_s^{(t)}\big)\Big\}.$$

Association is summarized by the posterior distribution of the correlation
$$\rho_{\text{seq}}^{(k,t)} = \operatorname{cor}\Big(Q_s^{(k)},\, \chi_s^{(t)}\Big),$$

computed across the 64 sequences. Posterior medians, 95% credible intervals, and $P(\rho_{\text{seq}} > 0)$ are reported for each rule k.

In addition to these aggregate summaries, sequence-specific correspondence is described by reporting, for each sequence s and rule k, the posterior probability

$$P\Big(\operatorname{sign}(Q_s^{(k)}) = \operatorname{sign}(\chi_s)\Big),$$

which highlights sequences for which a given canonical rule aligns closely with, or systematically diverges from, the anchor-based classification. These sequence-level diagnostics are used descriptively to assess heterogeneity in rule performance across the sequence space.

Participant-level similarity. Analogously, for each rule k, trial-level classifications are aggregated across sequences to form participant-level rule-based scores

$$Q_i^{(k)} = \frac{1}{N_i^{(k)}} \sum_s q_{is}^{(k)},$$

where $N_i^{(k)}$ denotes the number of included betting trials for participant i under rule k.

Let $\chi_i^{(t)}$ denote posterior draws of the participant-level directional tendency (the Gamblerish–Hot-handish Index) from EX1. For each rule k and posterior draw t, aggregate directional agreement is computed as

$$\text{Agree}_{\text{part}}^{(k,t)} = \frac{1}{N} \sum_i \mathbf{1}\Big\{\operatorname{sign}(Q_i^{(k)}) = \operatorname{sign}(\chi_i^{(t)})\Big\}.$$

Association is summarized by the posterior distribution of the correlation

$$\rho_{\text{part}}^{(k,t)} = \operatorname{cor}\Big(Q_i^{(k)}, \chi_i^{(t)}\Big),$$

computed across participants. Posterior medians, 95% credible intervals, and $P(\rho_{\text{part}} > 0)$ are reported for each rule k.

A Bayesian Modelling Framework and Robustness Procedures

All Bayesian models are estimated in Stan using dynamic Hamiltonian Monte Carlo. Unless stated otherwise, models use four chains and at least 2 000 post–warmup draws per chain. Convergence requires $\hat{R} \leq 1.01$ and an effective sample size of at least 400 per parameter.

A.1 Priors

Hierarchical models use weakly informative priors unless specified otherwise. Regression coefficients and location parameters follow $\mathcal{N}(0, 1)$; scale parameters follow HalfNormal$(0, 1)$. For proportion-scaled outcomes, $\mathcal{N}(0, 0.2)$ and HalfNormal$(0, 0.2)$ are used.

A.2 Likelihoods

Likelihoods follow standard formulations with structured replacement rules. For binary outcomes, a Beta–Binomial likelihood replaces the Bernoulli model when posterior predictive variance deviates substantially from observed variance. For non-negative continuous outcomes (such as welfare losses in RQ3), the primary likelihood is a hurdle–Gamma model, which accommodates exact zeros when they occur and flexibly models positive right-skewed values. A Gaussian working model with the same hierarchical structure may also be fitted as a simplified diagnostic alternative. For bounded outcomes with clustering at the upper boundary, a one-inflated Beta–Binomial model is used.

A.3 Posteriors

Whenever derived quantities depend on participant-level parameters obtained from auxiliary tasks, these quantities are recomputed at every MCMC iteration so that posterior uncertainty is propagated throughout all subsequent models. Posterior predictive checks compare observed and simulated distributions of key outcomes, and all estimation follows standard Stan and Bayesian workflow procedures.

A.4 Session Drift Adjustment

To accommodate potential within-session changes such as learning or fatigue, models may include a drift term when diagnostics indicate systematic temporal patterns. The experiment consists of four equal blocks (1–4).

Categorical drift uses

$$\gamma_{\text{block},k} \sim \mathcal{N}(0, 0.3), \qquad k = 1, \ldots, 4,$$

subject to the identifying constraint $\sum_{k=1}^{4} \gamma_{\text{block},k} = 0$.

Linear drift uses the centered index

$$\widetilde{\text{block}} \in \{-1.5, -0.5, 0.5, 1.5\}, \qquad \gamma_{\text{drift}} \sim \mathcal{N}(0, 0.3).$$

The drift component enters additively and is included only when justified by diagnostics.

A.5 Sensitivity and Robustness Analyses

Sensitivity analyses assess dependence on prior choices, likelihood specifications, and practical thresholds. Thresholds (e.g., ρ) are re-evaluated at levels 40% above and below their nominal values. Weakly informative priors are broadened to $\mathcal{N}(0, 2)$ or HalfNormal$(0, 2)$ where relevant. Alternative likelihoods from the replacement rules above are fitted when posterior predictive checks indicate misfit.

Substantive conclusions rely on the stability of posterior distributions across these alternative specifications rather than on any single model formulation.

A.6 Parameters and Indices

A.6.1 Holt–Laury Risk Preferences (r)

Individual risk preferences r_i are estimated from the Holt–Laury (HL) task using a hierarchical Bayesian model in Stan. Choices follow a CRRA utility function with individual-level parameters for risk aversion r_i and choice precision λ_i. Group-level priors follow Bland (2023)[1],

$$r \sim \mathcal{N}(0.27,\, 0.36), \qquad \lambda \sim \text{Lognormal}(\log 30,\, 0.5).$$

Expected-utility differences between the two HL lotteries enter a logistic choice rule. Posterior draws $r_i^{(m)}$ are used in all subsequent analyses; in every MCMC iteration, all quantities depending on r_i are recomputed so that uncertainty in risk preferences is fully propagated.

A.6.2 Expected-Utility–Optimal Stakes (a, a^*)

Preferences follow the CRRA utility function

$$u(x; r) = \begin{cases} \dfrac{x^{1-r}}{1-r}, & r \neq 1, \\ \ln x, & r = 1, \end{cases} \qquad x > 0.$$

All utility evaluations require strictly positive wealth. Whenever a stake would yield non-positive wealth in the losing state (e.g., $e - a \leq 0$ when $a = e$), that payoff is replaced by a small positive constant $x_{\min} = 0.01$ ECU when computing $EU(a; r)$.

Expected utility for stake a from endowment $e = 100$ with multiplier $m > 1$ is

$$EU(a; r) = \tfrac{1}{2}\, u(e - a + ma; r) + \tfrac{1}{2}\, u(e - a; r).$$

The expected-utility–optimal stake is defined as

$$a^*(r) = \arg \max_{a \in \{0,1,\ldots,e\}} EU(a; r),$$

with ties broken in favour of the smaller stake. This discrete definition applies uniformly to all r and ensures correctness for risk-loving ($r < 0$), risk-neutral ($r = 0$), and risk-averse ($r > 0$) preferences.

For $r > 0$ and $r \neq 1$, the interior optimum has the closed form

$$a^\star(r) = e\, \frac{(m-1)^{1/r} - 1}{(m-1)^{1/r} + (m-1)},$$

[1] Bland, James R., Bayesian Model Selection and Prior Calibration for Structural Models in Economic Experiments: Some Guidance for the Practitioner (January 24, 2023). Available at SSRN: https://ssrn.com/abstract=4334267 or http://dx.doi.org/10.2139/ssrn.4334267

and for $r = 1$,

$$a^\star(1) = e\,\frac{m-2}{2(m-1)}.$$

These expressions are used only when numerically stable and within $[0, e]$; the resulting value is then rounded to the nearest integer ECU. For $r \le 0$ or whenever the closed form is unstable (including $|r| < 10^{-6}$ and $|r-1| < 10^{-6}$), $a^*(r)$ is obtained directly from the discrete maximisation above.

Handling of extreme risk preferences. For a given multiplier m, the optimal stake $a^*(r_i; m)$ may be zero or strictly positive depending on individual risk preferences. In such cases, refraining from betting or engaging in betting, respectively, is normatively optimal at that multiplier.

For each participant i, we compute the posterior probability

$$P_i^0(m) = P\big(a^*(r_i; m) = 0 \mid \text{HL data}\big),$$

where $a^*(r_i; m)$ is computed using the discrete maximisation rule over $a \in \{0, 1, \ldots, e\}$ defined above. This probability is evaluated using posterior draws of r_i from the Holt–Laury model. Let $P_i^+(m) = 1 - P_i^0(m)$ denote the probability that the optimal stake is strictly positive.

Participants with $P_i^0(2.5) \ge 0.90$ are classified as *normative non-betters* and are excluded from confirmatory analyses that use a "betting is optimal" benchmark. Robustness checks additionally report results using alternative thresholds $P_i^0(2.5) \in \{0.80, 0.95\}$.

Participants with $P_i^+(1.9) \ge 0.90$ are classified as *normative betters*. Because analyses at this multiplier are exploratory, these participants are not excluded by default; results are additionally reported with and without this flagged subgroup for transparency.

A.6.3 Certainty Equivalent (c)

The certainty equivalent associated with a stake a is defined implicitly by

$$u\big(c(a; r); r\big) = EU(a; r), \qquad c(a; r) > 0.$$

For $r \ne 1$, this yields the closed form

$$c(a; r) = \big((1-r)\,EU(a; r)\big)^{1/(1-r)},$$

and for $r = 1$,

$$c(a; 1) = \exp\big(EU(a; 1)\big).$$

CRRA utility is defined only for strictly positive wealth. Whenever a stake would produce non-positive wealth in a payoff state (e.g., $e - a \le 0$ when $a = e$), that payoff is replaced by a small positive constant $x_{\min} = 0.01$ ECU when computing $EU(a; r)$. This ensures that both $EU(a; r)$ and $c(a; r)$ remain well defined for all posterior draws of r.

In practice, the closed-form expressions above are used whenever $(1-r)\,EU(a; r)$ is numerically stable and strictly positive. In any trial or posterior draw where the closed form is unstable (including regions where r is close to 0 or 1), the certainty equivalent is computed numerically as the unique $c > 0$ solving $u(c; r) = EU(a; r)$.

B Experimental Materials

B.1 Sequence Betting Task Decision Interface

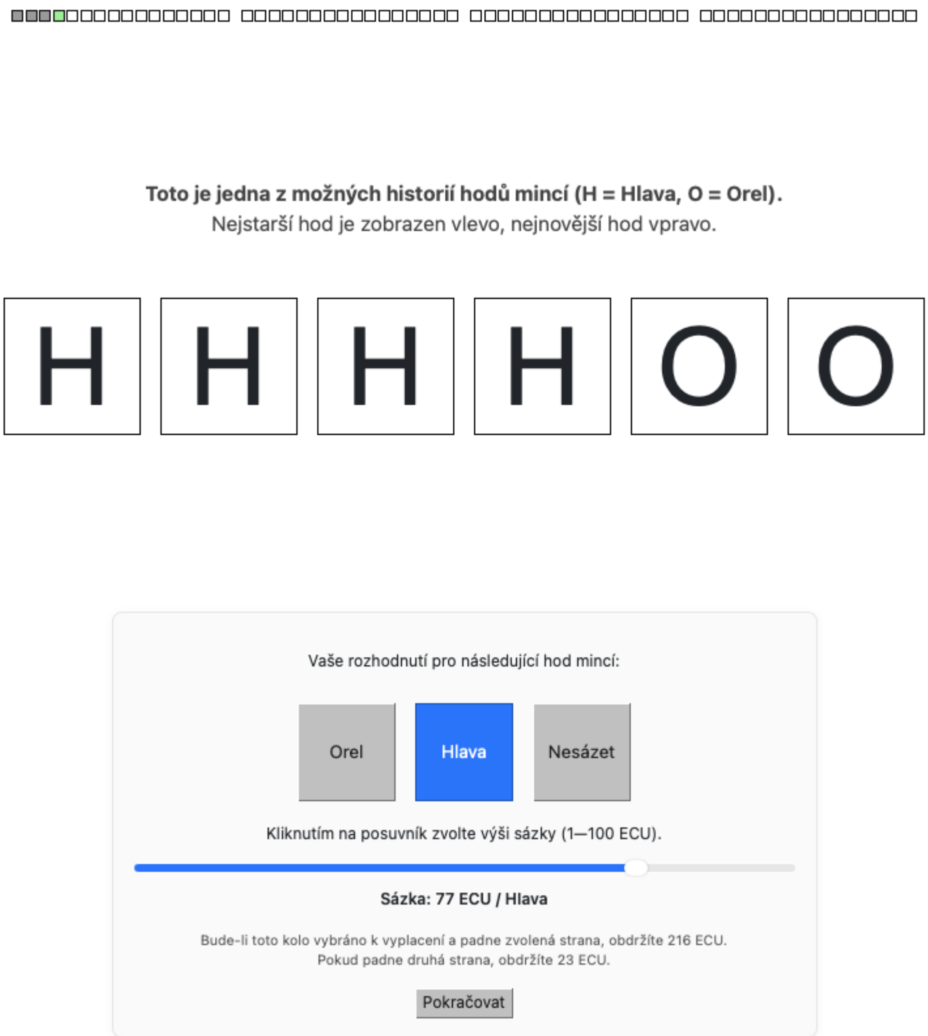

Figure 1: Decision screen presented in each trial of the main laboratory experiment.

Figure 1 illustrates the decision screen shown to participants in each round of the experiment. The top of the screen includes a progress indicator showing the participant's current position within the task (round and block progress). Participants observe a six-toss history of coin outcomes (H = Heads, O = Tails), displayed from left (oldest outcome) to right (most recent outcome). Below the sequence, participants choose whether to place a bet on the next coin toss by selecting one of three options: Heads, Tails, or No bet. If a betting option is selected, participants choose a stake using a slider ranging from 1 to 100 experimental currency units (ECU); choosing No bet corresponds to a stake of zero. The screen dynamically displays the implied payoff for the selected stake and side, conditional on the realization of the next coin toss. Participants confirm their decision by clicking the Continue button. The experimental interface is presented in Czech.