A Closer Look at the Gambler's Fallacy*

Lucas de Lara
Department of Economics, Columbia University
lpd2122@columbia.edu

November 14, 2024

Abstract

A classic explanation for the gambler's fallacy is that subjects believe that sequences of tosses from a fair coin should be representative of the randomness of the uniform distribution, and so should not have observable patterns. I introduce an information-theoretic formalization of this representativeness heuristic in terms of complexity and contrast it to the existing recency-weighted reversal model by Rabin and Vayanos. In order to test between these explanations, I collect rich choice and belief data from subjects predicting the next item from fully randomized sequences of binary outcomes, allowing me to take the analysis to the level of individual sequences. The basic results confirm the existence of the gambler's fallacy in the aggregate. However, there is also significant heterogeneity among subjects. I identify four types, depending on whether they report correct beliefs or incorrect beliefs that go in the gambler's fallacy direction, its opposite or a mix of both. Taking this heterogeneity into account, both models perform well when looking at an aggregate level, but a closer look at individual sequences reveals violations of the representativeness model which lead to a superior performance of the recency-weighted reversal model. The main component of this superior performance comes from the recency bias that subjects exhibit, which is not accounted for in the representativeness model.

JEL Classifications:
Keywords: Gambler's Fallacy, Representativeness, Complexity, Empirical Entropy

---

*Click here for lastest version.

I thank the members of the Cognition and Decision Laboratory at Columbia University and I recognize the financial support provided by the Columbia Experimental Laboratory for Social Sciences (CELSS).

1 Introduction

Imagine yourself observing someone toss a fair coin ten times in front of you, resulting in the sequence: head, head, head, head, head, head, head, head, head, head. This outcome was just as statistically likely as any of the other ones that could have happened, but a substantial and long-standing body of literature indicates that such an occurrence challenges commonly held but incorrect intuitions about randomness. The *gambler's fallacy* refers to a bias from which individuals erroneously anticipate too many reversals in sequences of independent random events. In the case mentioned, some might believe that, if the coin is truly fair and were to be tossed again, the next outcome would be more likely to – or even would *have to* – result in a tail.

The gambler's fallacy has long received learned attention. Oskarsson et al. (2009) provide a detailed survey of the interest and investigation it has received from several fields. Documentation of the phenomenon goes back at least as far back as the late eighteenth century, with Laplace noting that people would bet on numbers that they expected to be "due" for a draw in the French lottery, believing that "because the number has not been drawn for a long time, it, rather than the others, ought to be drawn on the next draw." (Dale and Laplace 1825/1995, p.92) His observations highlight a persistent tendency to misinterpret independent random events, expecting outcomes to self-correct in the short term to align with long-term statistical properties. This expectation of self-correction can even apply to one's own behavior. In the Zenith radio experiments in telepathy, happening in the late 1930s, listeners were invited to send their predictions over sequences of binary outcomes. Skinner (1942) was the first to point out that people's guesses were sequentially autocorrelated and exhibited what was to be known as the gambler's fallacy effect, first bringing it to the attention of psychologists. Since then, a large literature has developed in pursuit of explanations for how and why people hold these erroneous beliefs.

Despite the substantial size of this literature, existing theories remain lacking in either formalization or complete empirical testing. In this paper, I formalize a classical theory of the gambler's fallacy and compare it to the most recent model in the literature. In order to do so, I conducted an experiment that provided the data richness required to properly compare these two theories. Kahneman and Tversky (1972) attributed the gambler's fallacy to the representativeness heuristic. I return to a close reading of their description of this application, which reveals a connection between representativeness in the case of a fair coin and the complexity of the sequences of outcomes that it generates. I then develop a formalization which closely follows their words. This formalization brings tools from information theory to characterize randomness as information complexity, measured by the (higher order) empirical entropy of sequences. This measure includes moments of higher order than the basic proportions that have been the focus of existing models. In the most recent of such models, Rabin and Vayanos (2010) model the gambler's fallacy using two parameters, a reversal parameter which captures the proportion in previous outcomes, and a discount parameter which captures a recency bias effect. This model, however, still lacks a detailed empirical analysis, especially one that can differentiate between the two effects created by the two parameters.

Conducting the appropriate empirical analysis of these two theoretical explanations and comparing their performance requires a richness of data that has been missing in the literature thus far. Previous experiments have often used tasks that are difficult to incentivize or analyze, have often narrowly focused on a limited or preselected set of sequences of outcomes and have utilized limited data collection methods. By running an online experiment in which I collect, from a broader population, comprehensive choice and belief data for all possible sequences, under randomization, I obtain a richer dataset that enables detailed sequence-by-sequence analysis which allows for differentiating between competing explanations. The data confirms classical results regarding the choice of the next outcome when faced with full runs of the same outcome. However, closer inspection reveals important heterogeneities across subjects. Some subjects behave rationally, giving the correct answer, while a similar number indeed follows the gambler's fallacy, and yet others, in smaller numbers, actually behave in the opposite way. Focusing on the appropriate types of subjects, both the Rabin and Vayanos model and the entropy analysis perform well at the aggregate level, with highly significant estimated parameters. Taking the analysis to the level of individual sequences reveals the differences between the two: the Rabin and Vayanos model captures very well the observations in the data, while for some sequences, subjects' behavior contradicts the predictions coming from entropy.

The conceptual move to make the gambler's fallacy a case of a broader feature of cognition was made by Kahneman and Tversky (1972). They considered the gambler's fallacy to be a specific instantiation of the representativeness heuristic, wherein individuals expect even small samples to reflect the characteristics of the larger population or process from which they are drawn.[1] When applying the representativeness heuristic to the case of the gambler's fallacy, Kahneman and Tversky included in their definition of representativeness a component relating to people's intuitions about randomness, which is summed as a lack of regularity, a component that has escaped a proper formalization in existing models. Based on this notion of randomness, they proceed to say that, in the case of outcomes from a fair coin, they "venture that only HTTHTH appears really random. For four tosses, there may not be any" (ibid., p. 436). The cautious, non-committal language, though, highlights the likely absence of any formal and specific criteria for such judgements. Nonetheless, as I will show, a close reading of their verbal descriptions reveals links between their interpretation of randomness and notions of data compression, familiar from Information Theory Shannon (1948). Apparent randomness is linked to complexity, an absence of patterns

[1]Since the seminal paper, the representativeness heuristic has gone on to produce an expansive literature of its own. See, for example, Tversky and Kahneman (1974), Tversky and Kahneman (1982), Bar-Hillel (1984). For the purposes of this paper, however, the most important points to notice regarding this literature are the following. The main application of the heuristic has been to judgements of categorization (how likely an object is to belong to a certain set/category). The application to the gambler's fallacy has received comparatively little attention (it receives, e.g., passing mention in Bar-Hillel (1980, p. 588) and Bar-Hillel (1984, p. 103)). One kind of formalization that has existed in the literature is regarding base-rate neglect in Bayesian inference (e.g. Grether 1980), but this cannot explain the gambler's fallacy as it goes in the opposite direction of inference by evidence. The closest formalization to the one being presented here comes from Tenenbaum and Griffiths (2001), which formalizes representativeness as relative model evidence in the context of Bayesian inference. This approach has some important similarities with my formalization, to be discussed when it is introduced in subsection 2.2, but will also be shown to have three important differences: (i) it is not a model of belief formation and prediction, and is more appropriate to what I call in section 3 "randomness judgement tasks", which is indeed the type of task that their experiment on coin flips is; (ii) it lacks the connection to information theory and complexity that I establish; (iii) it leaves open what is the appropriate space of alternative hypotheses/models, whereas it will be shown that the complexity connection automatically establishes a relevant space of models.

in observed outcomes, which represents incompressibility of information.

I then follow this interpretation and formalize the gambler's fallacy as an expectation that outcomes from a fair coin will maintain a high level of information complexity in the sequence of outcomes that they produce. This complexity is measured by empirical entropy (Manzini 2001; Ferragina and Manzini 2005), which measures the proportions of outcomes coming after various combinations of previous outcomes. This means that sequences with low empirical entropy are those with clearer patterns, which allow for greater information compression, while high empirical entropy implies a lack of such patterns, lower availability for compression, and therefore high complexity. The connection to randomness and representativeness of the uniform distribution of a fair coin comes from the equivalence of empirical entropy and likelihood inference within the space of Markov chains (Gagie 2006). Sequences with high complexity are those for which the best model fit comes closer to the uniform distribution, which is the case of maximum entropy.

I contrast this formalized information theoretic version of the representativeness heuristic to Rabin (2002) and Rabin and Vayanos (2010). In the first one, the gambler's fallacy is modeled as a belief that outcomes are being drawn without replacement from a finite number of possible outcomes. The second, and more recent one, is a recency-weighted reversal model in which the gambler's fallacy is the joint result of two parameters: a reversal parameter, which captures the first moment of proportions and the expectation that this proportion will tend towards a balanced one, and a discount factor, which captures a recency bias in that more recent outcomes are weighted more heavily than those further past.

The experimental literature, in turn, has focused on the common paradigmatic tasks of choosing future outcomes after seeing a given sequence of outcomes, constructing random sequences and judging the randomness of given sequences. A common issue shared by all these tasks, similarly observed by Rabin (2002, p. 782) and repeated by Rabin and Vayanos (2010, fn. 10 and 11), is that, beyond just often not having been incentivized in existing studies, they are hard to incentivize at all. For all of them, there are no right or wrong answers, and it is hard to argue conclusively that there is anything wrong with any behavior observed from subjects. This is so even if specific and suggestive patterns are found in the data. The most natural solution is to focus on beliefs. In the case of a fair coin, the only correct belief to hold over the next outcome is the uniform Bernoulli distribution, which matches the true known data generating process. I leverage this fact to design a properly incentivized task. Nonetheless, it's still useful to collect choice data as well to keep within the tradition of the literature, and so I collect both choice and belief data, in this order and required to be consistent, for each sequence.

In contrast to most of the literature, Benjamin et al. (2017) made significant advances with a series of well-incentivized experiments. However, a significant limitation in their approach is that they only asked for belief responses for sequences of full runs of a single outcome. For other sequences, choice data was collected. This means that they could estimate the Rabin and Vayanos model with sequences of the former type, but could not precisely test it against those of the latter type. Moreover, they could not differentiate between the two effects present in the model, the reversal effect and the recency effect. This paper, on the other hand, allows for a detailed test of the model, and results show that the recency effect, in particular, plays a major role in explaining the fitting of the model to the observed data.

To obtain the detailed data necessary for the analysis mentioned above, I designed and ran an online experiment which proceeded as follows. Subjects were faced with a series of sequences of outcomes from flipping a computerized fair coin, heads and tails, and were asked to make two responses on each sequence. First, they had to choose by clicking on which outcome, head or tail, they wanted to bet on being the next one to happen. Second, they had to move sliders to give, in integer numbers from 0 to 100, their probability prediction for the likelihood of the next outcome being either a head or a tail. The sliders were programmed so that their probability response had to be coherent with their previous choice response, that is, if the subject chose head as the next outcome, their probability response for head had to be at least 50. This helps linking choice and probability responses as well as making choice decisions more meaningful for subjects which really believe that the probabilities are not 50-50.

297 subjects were recruited on the Prolific platform. Each subject faced 54 sequences. The first 50 were fully randomized, and consisted of 2 sequences of length 0, 4 sequences of length 2, 12 sequences of length 4, 20 sequences of length 6, and 12 sequences of length 8. The final 4 rounds consisted, in random order, of 1 sequence of each non-zero length but of the full run type. At the end, subjects answered a final non-incentivized follow-up multiple choice question about their reasoning during the experiment.

Plotting choice and probability response data for the full run sequences confirms a gambler's fallacy effect, especially on choice data as has been classically done in the literature. But a weaker effect on probability as well as a weaker monotonicity on the length of the run than might be expected suggest a more complicated picture. The first major result that departs from the previous literature is the identification of significant heterogeneity among subject types. Based on their probability responses to the last 4 rounds, subjects were classified into four types:

1. *rational*: Subjects who responded 50-50 probabilities in all rounds. Their responses demonstrate an understanding of the behavior of a fair coin and probabilistic independence.

2. *gambler's fallacy*: Subjects whose responses were consistent with the gambler's fallacy, expecting reversals of runs to be more likely.

3. *hot hands*: Subjects who showed the opposite pattern to the above, expecting runs to be more likely to continue than to break.

4. *both/mixed*: Subjects showing inconsistent patterns, sometimes predicting reversals and other times continuations.

The respective proportions of these types were 29.62%, 28.61%, 19.52% and 22.22%. Comparing subjects' types to their responses in the follow-up question revealed significant consistencies between the two. 'rational' subjects overwhelmingly respond that past flips do not affect future flips. 'gambler's fallacy' subjects overwhelmingly gave the expected response that the coin should generate balanced outcomes and therefore reversals were more likely. 'hot hands' subjects had more diversity in their responses, but they were consistent in giving responses that logically explain their answers. Then, looking separately by type at probability and choices responses for full runs confirms the patterns that one might expect from this classification, with significant effects in the appropriate direction for 'gambler's fallacy' and 'hot hands' subjects as well as monotonicity.

Given subject heterogeneity, the proceeding analysis was done separately by subject type, with a special focus on 'gambler's fallacy' subjects. The first step is to fit both the Rabin and Vayanos model and the empirical entropies model to probability responses for all sequences. This gives highly significant parameters in both cases. In the Rabin and Vayanos case, both reversal and recency parameters are highly significant and of sizes that produce substantial effects on both fronts. For empirical entropies, the parameters for zeroth, first and second order entropies are all significant, but their parameter sizes are decreasing, showing that more complex patterns have less impact. Both perspectives, therefore, confirm at an aggregate level the impact of sequence properties beyond the basic proportion between the two outcomes. But a closer analysis is required to differentiate between the two.

The detailed nature of the data collected is then leveraged to properly compare the two models. The analysis is taken to the level of individual sequences of length up to 6 to contrast mean predictions against the predictions of each model. This level of analysis reveals that the Rabin and Vayanos model performs significantly better than predictions coming from sequence complexity. In particular, there are sequences, most notably the alternating sequences like *head-tail-head-tail...* for which subjects predict a continuation of the pattern rather than a break. This goes against the representativeness model, as a pattern break increases complexity. It is aligned, however, to the recency effect in Rabin and Vayanos model, as one of the outcomes is further to the front of the sequence than the other, and is thus more heavily weighted in the expectation of reversal.

The analysis thus far finds that the Rabin and Vayanos model performs significantly well against the data. The final step of the analysis is to consider whether this model, which consists of two parameters can improved upon by a generalized model with eight parameters, one for each possible position in the sequences. The results reveal that the estimated generalized parameters are very close to those implied by the two parameters in the original model, confirming once again its excellent fit to the data.

One aspect to be highlighted is that the gambler's fallacy is particularly puzzling from a bounded rationality perspective. Especially in the context of this paper's experiment, the task at hand is simple and it's hard to see what cognitive constraints might be operant and leading to the wrong answers. If subjects have no sense of what outcomes from a fair coin are supposed to be, it would make sense for their answers to follow the pattern of statistical inference, expecting, for example, heads to be more likely if that's what they observe. This is indeed how the 'hot hands' subjects mentioned above behave. But the gambler's fallacy goes in the opposite direction. By studying this kind of behavior, which might be described as a kind of anti-inference, it might be possible to have a better understanding of the mechanisms of belief formation in general. For belief formation, and predictions generated by them, permeate several, if not all, aspects of economic decision making. This is particularly acute for domains where probabilistic thinking is most essential for reaching conclusions, such as financial decision making.

This paper is structured as follows. Section 2 provides some notation and the theoretical exposition of the two models being considered and compared. Section 3 introduces the experimental design and explains how it was implemented. And section 4 provides the results from analyzing the data collected from the experiment. A conclusion discusses the results.

2 Theory

2.1 Notation

An agent faces finite binary strings of varying length $n \in \mathbb{N}$. They are denoted $s^n \in 2^\omega$. For example, $s^4 = 0011$ or $s^6 = 101110$. The data generating process producing these strings is given by independent draws from a distribution $p \in \Delta(\{0,1\})$, and the agent is informed of this process. In the case of a fair coin, p is the uniform Bernoulli distribution. After seeing a sequence s^n, the agents reports a belief q over the next outcome. That is, the agent's responses follow a belief mapping $s^n \mapsto q \in \Delta(\{0,1\})$. Let q_0 be the probability given for outcome 0. In some cases, the affine transformation $x \mapsto 2x-1$ and its inverse are used, so that it may be that $q_0 \in [0,1]$ or $q_0 \in [-1,1]$.

In the gambler's fallacy, q deviates from p in a systematic way. For example, despite p being uniform, the agent might report $q_0(s^6) < 0.5$ if $s^6 = 010000$.

2.2 Complexity as Representation of Randomness

The gambler's fallacy can be modeled as people expecting that outcomes from a fair coin should not produce noticeable patterns. Instead, they expect that the sequence of outcomes is supposed to maintain a high degree of complexity, which will be shown to mean an absence of patterns of any kind. This complexity is linked to the kind of randomness that is expected from the uniform distribution. In this sense, sequences of higher complexity are more representative of this kind of randomness than patterned sequences. This subsection shows how this idea can be expressed formally via the complexity measure of empirical entropy in a way that captures the original notion of the representativeness heuristic.

Kahneman and Tversky (1972) introduced the idea of the representativeness heuristic and presented the gambler's fallacy as one of its applications. According to this notion of representativeness, the "subjective probability of an event, or sample, is determined by the degree to which it: (i) is similar in essential characteristics to its parent population; and (ii) reflects the salient feature of the process by which it is generated." In the case of a fair coin, the process is the binary uniform distribution, linking a representative sample to the kind of randomness that is expected from such a distribution. "A representative sample, then, is similar to the population in essential characteristics, and reflects randomness as people see it; that is, all its parts are representative and none is too regular." (ibid. p. 436)

A formalization of this idea requires further clarification regarding what it means for a sequence to appear random, and for its parts to not be 'too regular'. The following words, worth quoting in full, offer strong hints of how one should proceed:

> Random-appearing sequences are those whose verbal description is longest. Imagine yourself dictating a long sequence of binary symbols, say heads and tails. You will undoubtedly use shortcut expressions such as 'four Ts,' or 'H-T three times.' A sequence with many long runs allows shortcuts of the first type. A sequence with numerous short runs calls for shortcuts of the second type. *The run structure of a random-appearing sequence minimizes the availability of these shortcuts, and hence defies economical descriptions.* ***Apparent randomness, therefore, is a form of complexity of structure***. (ibid. pp. 436–437, both italics and bold mine)

Based on this, I turn to Information Theory to formalize random-appearing sequences as those with a high degree of complexity, which, equivalently, are those with low levels of information compressibility. Inversely, less complex sequences are those with patterns that can be exploited for information compression. This is captured via the complexity measure of (higher order) empirical entropy (Manzini 2001; Ferragina and Manzini 2005).

Let $\Sigma = \{\alpha_1, ..., \alpha_h\}$ be a finite alphabet. Let s^n be a string of n symbols from Σ, and let n_i, $i = 1, ..., h$, be the number of occurances of symbol α_i in s^n. The (zero-th order) empirical entropy of s^n is given by[2]

$$H_0(s^n) = -\sum_{i=1}^{h} \frac{n_i}{n} \log\left(\frac{n_i}{n}\right).$$

This is equivalent to the entropy of the empirical distribution of the sequence. $nH_0(s^n)$ measures the maximum compression of the sequence that's achievable via a uniquely decodable code with a fixed codeword for each symbol. Note that in the binary case, this simply registers the relative proportion of the two possible symbols. In the case of a fair coin, the relative proportion between heads and tails. It equals 0 if the

[2]The logarithm being base 2 and assuming $0 \log 0 = 0$.

outcomes are all heads or all tails, and 1 if they're exactly balanced. Therefore, at this zero-th order level of consideration, an expectation of high complexity simply means an expectation that 0s and 1s ought to stay in a balanced proportion.

To capture more complicated patterns that might exist in the sequence or might be expected as more outcomes happen, one needs to consider the possibility of using codewords that depend not only on the symbol itself, but also on the symbols preceding it. For each $w \in \Sigma^k = \Sigma \times \Sigma \times ... \times \Sigma$ and $\alpha_i \in \Sigma$, let $n_{w\alpha_i}$ denote the number of times that the substring $w\alpha_i$ appears in the string s^n, and let $n_w = \sum_i^h n_{w\alpha_i}$. The k-th order empirical entropy of s^n is given by

$$H_k(s^n) = \sum_{w \in \Sigma^k} \frac{n_w}{n} \left[-\sum_{i=1}^{h} \frac{n_{w\alpha_i}}{n_w} \log \left(\frac{n_{w\alpha_i}}{n_w} \right) \right],$$

with $H_{k+1}(s) \leq H_k(s)$ for any s and k. Understanding the formula is easier after knowing the procedure for calculating it. For each $w \in \Sigma^k$, one constructs the subsequence of symbols from s^n that come after w, then calculates the zero-th order empirical entropy of this subsequence (this is the value in the brackets), which is then weighted by n_w/n, and then moves on to the next element in Σ^k.

In the binary case, $0 \leq H_k(s) \leq 1$ for any k and s. Sequences with higher empirical entropy, closer to 1, are those to be considered more complex, up to the level of the order being considered. To quickly see how these higher orders differentiate the complexity of sequences with regard to patterns that go beyond basic proportions, consider the example of two sequences of length 8,

$$s^8 = 01010101 \text{ and } \tilde{s}^8 = 10100011.$$

Both have both outcomes in equal proportion, but the second one appears more complex than the first one, which actually seems highly patterned. And, indeed, we have

$$H_0(s^8) = H_0(\tilde{s}^8) = 1, \text{ but } H_1(s^8) = 0 \text{ while } H_1(\tilde{s}_8^2) \approx 0.8443.$$

While empirical entropy allows for the measuring of complexity of sequences, it's not yet clear that it provides any more rigorous justification for the notion that more complex sequences are those that appear more random, and thus more representative of the uniform distribution. The missing link is provided by the following result.

Theorem 1 (Gagie, 2006)**.**

$$H_k(s^n) = \frac{1}{n} \min_{P \in \mathcal{P}_k} \log \frac{1}{P(s^n)}$$

where $\mathcal{P}_k$ is the set of k-th order Markov chains.

The theorem states the empirical entropy of order k of a sequence s^n, as calculated by the formulas above, coincides with the (averaged by size) log-likelihood of the best explanation for the sequence within the space of k-th order Markov chains.

This shows a direct connection between the complexity of a sequence, measured by empirical entropy, and the best likelihood inference that its outcomes suggest within the space of Markov chains. Notice that the uniform distribution is always in the set $\mathcal{P}_k$. In the binary case, plugging in the uniform Bernoulli results, for any s^n, in $\frac{1}{n} \log \frac{1}{P(s^n)} = 1$, which is the level of maximum entropy. High complexity, then, is a worst case scenario in likelihood inference, in which no better explanation than the uniform distribution can be suggested by the outcomes of the sequence. This formalizes representativeness in terms of statistical inference while also connecting it to the idea of high complexity in the case of the uniform distribution.

The gambler's fallacy can then be modeled as an expectation that the next outcome in the sequence should be the one that makes the sequence more complex than the other. Given the development in the previous paragraphs, this means that the outcomes also makes the sequence more representative of the uniform distribution. To capture the difference that one outcome makes relative to the next, the following variables are introduced:

$$\Delta_0 H_k(s^n) = H_k(s^n 0) - H_k(s^n 1),$$

in which $s^n 0$ is the sequence s^n with a 0 added at the end, and likewise for $s^n 1$. The agent's belief q_0 over the next outcome being 0 becomes a function of these differences in empirical entropies $\Delta_0 H_k$.

The following table shows the example of the sequence $s^6 = 000111$, for which $\Delta_0 H_0(s^6) = 0$, $\Delta_0 H_1(s^6) > 0$ and $\Delta_0 H_2(s^6) > 0$.

Sequence	**Future Sequence**	H_0	H_1	H_2
000111	0001110	0.985228	0.787111	0.571429
	0001111	0.985228	0.393555	0.285714

I take an agnostic position with regards to the precise functional form that relates these variables $\Delta_0 H_k$ to beliefs. Given the theory that the gambler's fallacy consists in believing that complexity-increasing outcomes should be more likely, this means that beliefs will follow a function

$$q_0(s^n) = q_0(\Delta_0 H_0(s^n), \Delta_0 H_1(s^n), \Delta_0 H_2(s^n))$$

which is increasing in its arguments, and such that $q_0(0, 0, 0) = 0.5$.[3] In the example above of $s^6 = 000111$, the prediction is that $q_0(000111) > 0.5$.[4]

Below is another example, this time the case of $s^6 = 010101$, which has $\Delta_0 H_0(s^6) =$

[3]Given the binary alphabet being worked with, and that in the experiment sequences will be of length up to 8, only empirical entropies of up to order 2 will be considered, following the results from Gagie (2006)

[4]As discussed in footnote 1 in the introduction, this approach to representativeness has some similarities but mostly differences to that seen in Tenenbaum and Griffiths (2001). They present a measure of representativeness in terms of relative model/hypothesis evidence in terms of Bayesian inference. How representative observed data d is of a model/hypothesis $h_i \in \mathcal{H}$ is given by

$$R(d, h_i) = \log \frac{P(d|h_i)}{\sum_{h_j \in \mathcal{H}} P(d|h_j) P(h_j|\overline{h}_i)}$$

in which $P(h_j|\overline{h}_i)$ is the prior probability of h_j given that h_i is not the correct model/hypothesis. This approach is similar to mine in considering representativeness as *relative* representativeness, comparing likelihoods between different possible models. The first immediate difference is the Bayesian framework with the use of prior beliefs. The second is a lack of connection to information theory and complexity. The third is that this is not extended to belief formation and prediction, and their experiment indeed only asked subjects to judge how likely four preselected sequences were to be coming from a few alternative coins (fair, alternating, biased and deterministic). The fourth is that the space of possible models $\mathcal{H}$ being considered is left undetermined, whereas the complexity connection establishes a specific set of possible models, namely the set of Markov chains of a specific order. The fifth is that my approach compares the reference distribution, the uniform Bernoulli, to the *best* alternative explanation within this space of possible models.

0, $\Delta_0 H_1(s^6) < 0$ and $\Delta_0 H_2(s^6) < 0$. In this case, the prediction is that $q_0(010101) < 0.5$.

Sequence	**Future Sequence**	H_0	H_1	H_2
010101	0101010	0.985228	0.000000	0.000000
	0101011	0.985228	0.393555	0.393555

2.3 Reversal and Recency

The formalization of the representativeness heuristic in the previous subsection was capable of capturing the effects of higher order complexity. This subsection, in turn, presents the most recent model of the gambler's fallacy in the literature, the recency-weighted reversal model presented by Rabin and Vayanos (2010). Their model is also capable of capturing the effects on beliefs of more complicated patterns than simple proportions. It models the gambler's fallacy through the combination of two effects, each with its own parameter, a reversal effect and a recency effect. It's the latter one that is necessary for creating a gambler's fallacy effect that goes beyond the first moment of relative numbers of the two outcomes. This model makes predictions that sometimes go in the same and sometimes in the opposite direction of the predictions coming from the representativeness model, opening the way for differentiating between them using experimental data.

In the general version of the model, an agent needs to make predictions on future outcomes based on inferences they make about an underlying moving state. The agent receives noisy signals about the state, and the gambler's fallacy is modeled as the 'mistaken belief' that the noise shocks in the signals are not i.i.d. as they actually are, but rather must exhibit reversals with regard to previous outcomes, which are weighted by a discount factor.

In the case of an agent observing the results from tossing a fair coin, the underlying state can be taken to be fixed, or non-existent, leading to a belief function that takes the form[5]

[5]This is the same form, with adjusted notation, seen in Benjamin et al. (2017, p. 15).

$$q_0(s^n) = \alpha \sum_{t=0}^{\infty} \delta^{t+1} s^n_{n-t} \in [-1, 1]$$

where q_0 is probability of 0, and s^n_{n-t} codes the value in position $n-t$ of the sequence:

$$s^n_{n-t} = \begin{cases} 1, \text{ if } = 1, \\ 0, \text{ if } = \emptyset, \\ -1, \text{ if } = 0. \end{cases}$$

The parameters $(\alpha, \delta) \in [0, 1]^2$ capture, respectively, the reversal and recency effects. The reversal effect is stronger the closer α is to 1, while the recency effect is stronger the closer δ is to 0.

Notice that if $\delta = 1$, that is, if there's no recency effect, beliefs will be only a function of the number of excess outcomes in one or the other direction. In particular, any sequence with balanced numbers leads to a correct uniform belief, which applies to, for example, both examples from the previous subsection: 000111 and 010101.

It's when both $\alpha > 0$ and $0 < \delta < 1$ that both effects are activated and more complex predictions obtain from the model. It's immediate that, under such parameters, $q_0(000111) > 0.5$, which coincides with the representativeness prediction. For the sequence $s^6 = 010101$, since the 0s are further back than the 1s, the prediction is that $q_0(010101) > 0.5$, which goes in the opposite direction to the representativeness prediction.

The two theories presented in this section, then, make different predictions for certain sequences, opening up the possibility of testing and differentiating between them using experimental data.

3 Experiment

The experimental design had the main goal of producing a rich enough dataset that could be used to evaluate the performance of each one of the theories from the previous section as well as to compare them to each other. Beyond this, it also had the goal of improving upon previous experiments, which have mainly focused on three paradigms:[6]

i. Prediction Tasks: Participants are presented with a sequence of outcomes and asked to choose which outcome they think will happen next. This is the most straightforward test of the gambler's fallacy, but the analyses of these tasks have focused on choices after a streak of the same outcome, ignoring more complex patterns of interaction. Furthermore, the streaks are often short, such as two, three or four repetitions.

ii. Sequence Construction Tasks: Participants are instructed to generate sequences that they believe are random, in the sense of being most like what they would expect tosses of a fair coin to generate. These constructed sequences are then analyzed to check for frequencies of reversals and other statistical properties, and it's also possible to consider which sequences people avoid constructing.

iii. Randomness Judgment Tasks: Participants evaluate the randomness of given sequences, such as by being told that it may or may not have been generated by a fair coin and then being asked which case they think it is. Sequences can then be categorized according to such answers.

The main difficulty with tasks such as these is that there is not correct answer, since a fair coin is equally likely to produce any future outcome or any combination of outcomes. This complicates the issue of properly incentivizing responses, as well as judging whether such responses imply that subjects are truly wrong. The main way to fix this issue is to ask for *belief* responses. That is, how likely, in probability terms, subjects think each outcome is to come next. In this case there is one single objectively correct answer.

Having enough data richness to compare the two theories meant being capable of taking the analysis to the level of individual sequences. In order to achieve this, a large degree of randomization was employed in conjunction with a large number

[6] See Oskarsson et al. (2009, p. 264) for examples of each.

of tasks per subject. Subjects were presented with sequences of outcomes generated by a computerized version of a fair coin. These outcomes in these sequences were completely randomized, and their lengths varied from zero to eight outcomes, in increments of two, with the order of their appearance also being randomized. For each sequence, both choice and belief responses were collected.

3.1 Task

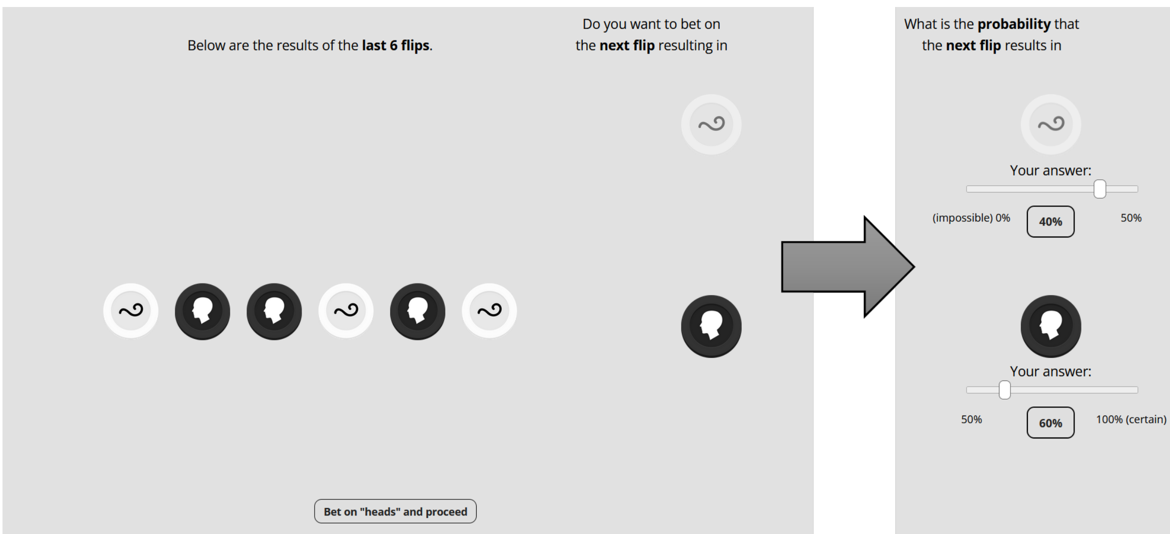


Figure 1: Typical task screen for a given round. Subjects are shown sequences of heads and tails and are asked for the next outcome and probabilities. The center-left of the screen shows the sequence of outcomes and remains the same throughout the round. The right side of the screen first asks for choice response, and then for probability responses. The button at the bottom of the screen is used to confirm both responses.

Figure 1 shows what a typical task screen looked like. In the center-left of the screen, a sequence of heads and tails was presented to the subject.[7] They then had to provide two responses by interacting with the right side of the screen and the confirmation button at the bottom. First, they had to click on which outcome, head or tail, they wanted to predict as being the next one. After that, sliders appeared

[7]A tail outcome was coded as 0, a head outcome was coded as 1. So, for example, the sequence appearing in figure 1 is 011010.

beneath each outcome. By moving the sliders, they recorded their probabilistic beliefs, in integers from 0 to 100, regarding the likelihood of each outcome being the next one. When one slider was moved, the other one adjusted automatically. The sliders were programmed such that the probability for the chosen outcome had to be at least 50%. After the responses, a new outcome was generated to check if it matched the prediction. No feedback was provided.

3.2 Structure

After consenting to participate, subjects were presented with instruction screens explaining the nature of the task to them and how their answers would affect their bonus payment. Subjects were explained that the computer was producing the sequences they would be seeing in a way that exactly mirrored a real fair coin. The instructions first presented an explanation of the choice response, which was followed by 6 'warm up' questions, in which subjects only had to choose the next outcome. Out of these 6 questions, one was chosen and if their prediction matched the actual result, $0.60 was added to their bonus payment.[8] Then the probability response was introduced and explained to them. At the end of the instructions, there was one practice question with both choice and probability responses just like a normal round. The bonus structure for choice and probabilities was explained to them and is detailed in the next subsection.

Subjects went through 50 rounds in which they encountered, in random order, 2 sequences of length 0[9], 4 sequences of length 2, 12 sequences of length 4, 20 sequences of length 6 and 12 sequences of length 8. In each of these sequences, each outcome was randomized, i.e. it was equally likely to be a head or a tail. After these 50 sequences, subjects encountered 4 more sequences, one of each of the previous lengths and in random order. Each one of these sequences was composed of either all heads or all

[8]This effectively amounted to an extra $0.30 fixed payment in expectation regardless of their actual choices and was only meant to fix their understanding of the task before introducing belief responses.

[9]The question language was similar but slightly changed from the one seen in figure 1 for length 0. Subjects were asked 'Do you want to bet on the **first flip** resulting in' and 'What is the **probability** that the **first flip** results in'.

tails, each case being equally likely. For compensation purposes, these last 4 sequences were treated just like the previous 50, and this was explained to subjects.[10]

Before finishing the experiment, subjects were asked one non-incentivized multi-choice followup question asking for their reasoning in responding to the full run sequences. They were presented with 6 possible answers. See figure 2.

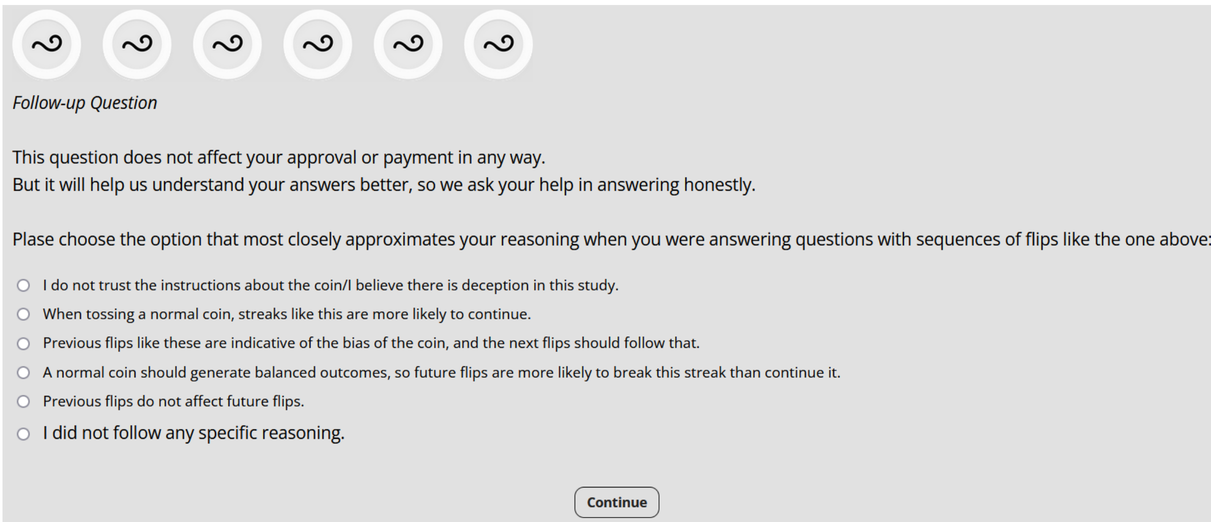


Figure 2: Follow-up question at the end of the experiment asking subjects for their reasoning for their answers when encountering full runs of heads or tails.

3.3 Implementation

297 subjects were recruited through the Prolific website, an online platform specialized in maintaining a pool of subjects for scientific research. Using this platform allowed for easy recruiting of a large number of participants from a broad population. Subjects were restricted to being located in the USA and being fluent in English.

Throughout the instructions, subjects were asked 4 simple comprehension questions about the functioning of the task and what the responses meant. In accordance with Prolific's policies on comprehension questions, subjects who answered the same

[10]Given that these last four sequences were not randomized in the same way as the previous 50, the language was adjusted to account for this. Subjects were asked what would their answers "have been" if they "had seen" these sequences.

question incorrectly twice were not permitted to continue. This happened to 18 subjects.[11]

The fixed payment for finishing the experiment was $2.70. The bonus payment varied from $0.00 to $7.00. A potential bonus of $0.60 came from the warm-up questions. Two questions out of the 54 were randomly selected for the remaining bonus. The choice response was worth a $1.00 bonus. The probability response was worth a $2.20 bonus, calculated according to the Binarized Scoring Rule (Hossain and Okui 2013). Following recent results from Danz et al. (2022), subjects were not directly given explicit mathematical formulas for this calculation, but were told that their bonus payment was maximized by truthfully reporting their best guesses for the correct probabilities. Mathematical formulas were provided at the end of the experiment.

The experiment was coded in JavaScript using custom plug-ins for the jsPsych library (de Leeuw 2015).

4 Results

In this section, I present the results of the analysis of the data from the experiment described in the previous section. I proceed from the aggregate to the granular level. I begin by showing that overall results from full runs of heads and tails look as one would expect from the existing literature. I then proceed to show, however, that there is significant heterogeneity among subjects, and they can be classified into four different groups: *rational*, *gambler's fallacy*, *hot hands* and *both/mixed*. The analysis then proceeds by treating these groups separately, focusing on subject types 'gf' and 'hh', as these are the types for which there are significant results.[12] The first step proceeds at an aggregate level to verify the overall performance of each of the two models presented in section 2. Both perform well, with highly significant parameters. Once the analysis is taken to the level of individual sequences, I show

[11] This is in addition to the 297 who successfully finished the experiment, resulting in a rate of 5.71%.

[12] Aggregate analysis for all subjects and the other subject types is provided in appendix B.

that the Rabin and Vayanos model performs better than the entropy theory. Given this superior performance, I then verify whether increasing the number of parameters in a generalized version of the Rabin and Vayanos model increases performance. The general parameters are very close to the implied parameters from the simple model and offer no increase in explanatory power.

4.1 Basic Results

Figure 3 follows the basic structure of several figures appearing in this section. It shows both mean probability (left, on a 0-100 scale) and choice frequency (right, on a 0-1 scale) data for a group of sequences. The sequences displayed are those of full runs of heads or tails for length 2, 4, 6 and 8.[13] The responses are for the outcome that continues the run, for example, frequency of choosing head as the next outcome when the sequence is a full run of heads. The responses are from all subjects.

Concentrating first on the choice data, it shows a significant gambler's fallacy effect. Subjects are significantly more likely to choose the outcome that breaks the run than the one that continues it. Probability responses are less definitive, but there's still a significant gambler's fallacy effect for lengths 4 and 6: subjects believe the outcome that continues the run is less likely to happen than the one that breaks it. As the next subsection will show, however, looking at all subjects in this way obscures important differences between them. Nonetheless, this basic result shows that choice data, at the aggregate level, might overestimate the gambler's fallacy effect. This further highlights the importance of the probability data collected in my experiment.

4.2 Subject Heterogeneity

In the last four rounds of the experiment, sequences subjects encountered always took the form of full runs of either heads or tails, one of each of the lengths 2, 4, 6 and 8.

[13] Except when noted, in these figures, symmetric sequences (e.g. 0110 and 1001) are pooled together.

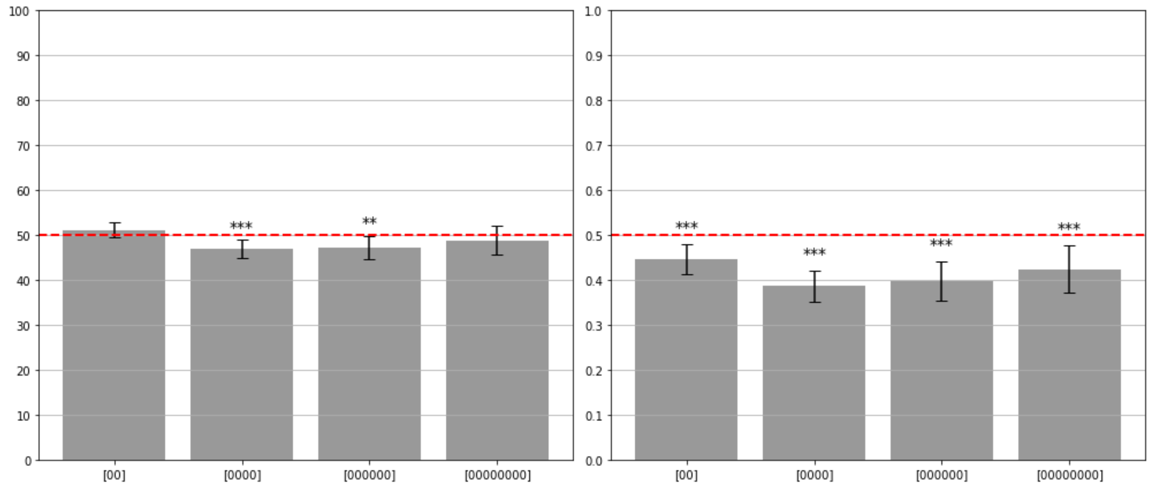


Figure 3: Mean probability (left) and choice frequency (right) responses for the outcome that continues the full run. All subjects. Error bars are 95% confidence intervals. Stars record level of significance at which the displayed values are different from 50 (left) or 0.5 (right): * for 10%, ** for 5%, *** for 1%.

Based on their probability responses for the outcome that continued the run in these rounds, subjects were classified into one of four types: *'rational (rat)'*, *'gambler's fallacy (gf)'*, *'hot hands (hh)'*, *'both'*. Table 1 shows the behavior that led to being classified into each one.

Type	**Classifying Behavior - Probability response for run-continuing outcome**
rat	50% on all four rounds
gf	weakly below 50% on all four rounds, strictly below on at least one round
hh	weakly above 50% on all four rounds, strictly above on at least one round
both	strictly above 50% on at least one round and strictly below 50% on at least one round

Table 1: How subjects were classified into each of the four types depending on their probability responses for the outcome that continued the full runs in the final four rounds.

Based on these criteria, table 2 shows the distribution of types. The most common type of subject was the *'rational'* type, encompassing almost 30% of subjects. It was closely followed by the *'gambler's fallacy'* type. The least common type was the *'hot hands'* type.

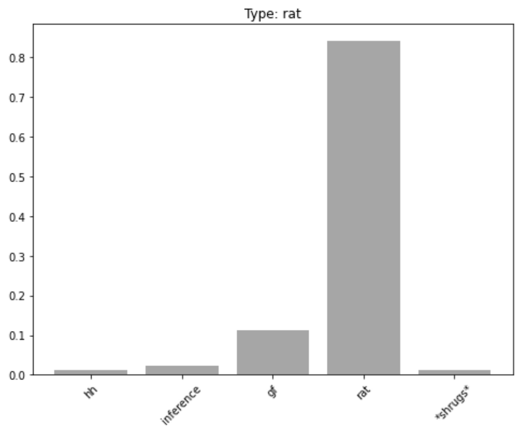
Type	Number	Proportion
rat	88	29.62%
gf	85	28.61%
both	66	22.22%
hh	58	19.52%

Table 2: Distribution of subject types.

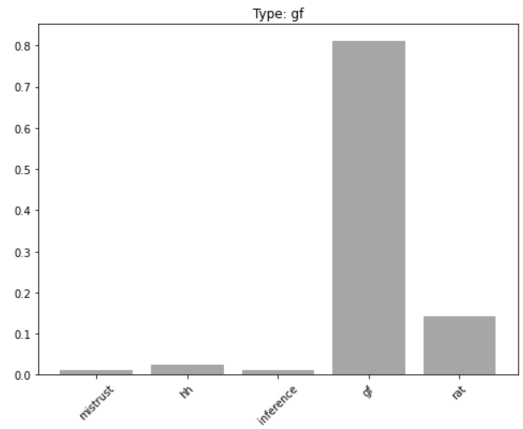
As explained in section 3.2 and shown in figure 2, at the end of the experiment subjects were asked to choose between 6 options that best matched the reasoning for their responses to full run sequences. These 6 options were, in order, labeled as follows: (i) "I do not trust the instructions about the coin/I believe there is deception in this study." as labeled 'mistrust'; (ii) "When tossing a normal coin, streaks like this are more likely to continue." was labeled 'hh'; (iii) "Previous flips like these are indicative of the bias of the coin, and the next flips should follow that." was labeled 'inference'; (iv) "A normal coin should generate balanced outcomes, so future flips are more likely to break this streak than continue it." was labeled 'gf'; (v) "Previous flips do not affect future flips." was labeled 'rat'; (vi) "I did not follow any specific reasoning." was labeled '*shrugs*'.

Figure 4 shows the distribution of these answer for each of the four types of subjects. The answers are mostly strongly in line with what would be expected, especially for 'rat' (figure 4a) and 'gf' (figure 4b) type subjects. About 85% and 80% of subjects of these types, respectively, gave the justification that matches their type. For 'hh' (figure 4c) type subjects, the three most common answers were those that could justify a belief that a full run of heads or tails is more likely to continue than to stop.

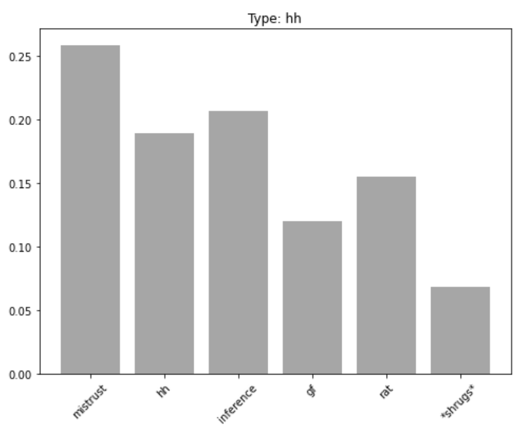
Figures 5, 6, 7 and 8 repeat figure 3 but considering individually each of the four types of subject: in order, 'gf', 'hh', 'rat' and 'both'. These figures show that the original aggregated data shown in the latter figure contained trends going in different directions. Types 'gf' (figure 5) and 'hh' (figure 6) show opposite tendencies: they expect the next outcome to be more likely to, respectively, reverse or continue the run, and this likelihood is monotone in the length of the run. Subjects of type 'rat'



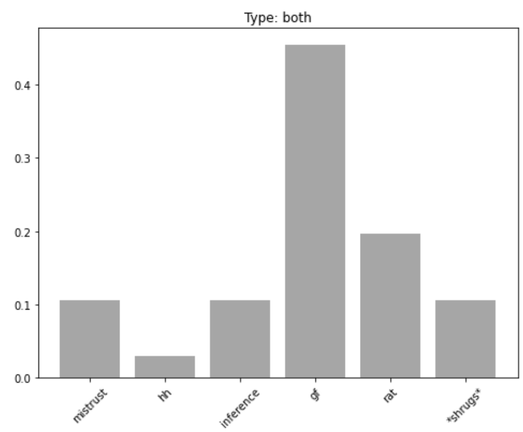
(a) Distribution of follow-up responses for *'rational'* type subjects.



(b) Distribution of follow-up responses for *'gambler's fallacy'* type subjects.



(c) Distribution of follow-up responses for *'hot hands'* type subjects.



(d) Distribution of follow-up responses for *'both'* type subjects.

Figure 4: Distribution of follow-up responses for different subject types.

(figure 7) display an interesting pattern: their beliefs match the correct distribution, but their choices exhibit a gambler's fallacy effect.[14] Naturally, it's not possible to affirmatively says that these subjects are 'wrong' in their choices for the reasons already discussed.[15] Finally, for subjects of type 'both' (figure 8), it's not really possible to offer any conclusive analysis. There's a significant gambler's fallacy effect

[14]Reminding the reader that both all-heads and all-tails sequences are included, so any systematic bias in the direction of one or the other outcome would be washed out by randomization.

[15]Given that the 50-50 response was always available, the restriction of choice and beliefs being consistent was non-binding for these subjects.

for the sequences of length 4, but this effects disappears for longer lengths, and it's not there for length 2 either. It's possible to speculate that these subjects might switch their behavior depending on the length of the run, but the data as it stands cannot be used to conclusively justify such a conjecture.

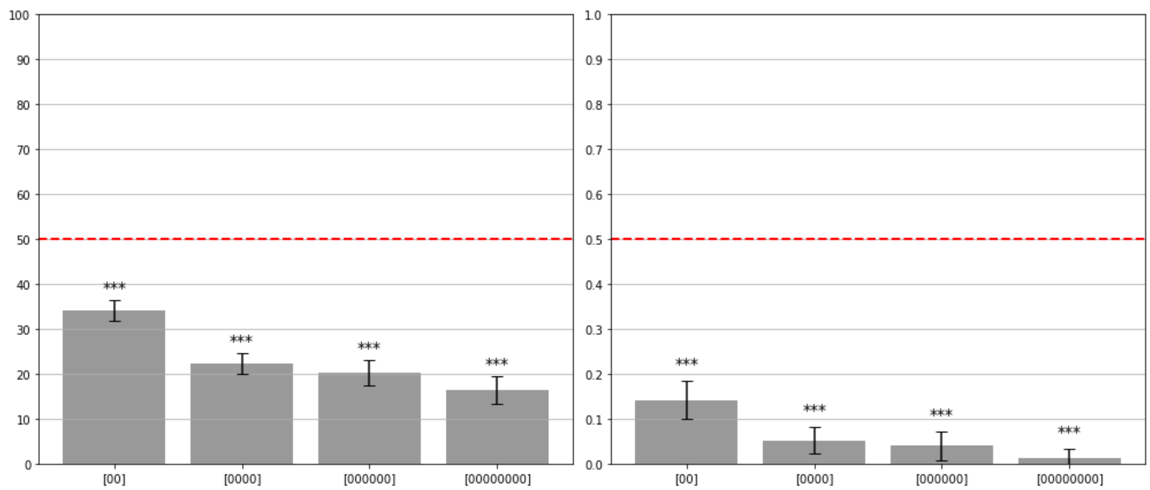


Figure 5: Mean probability (left) and choice frequency (right) responses for the outcome that continues the full run. 'Gambler's fallacy' subjects. Error bars are 95% confidence intervals. Stars record level of significance at which the displayed values are different from 50 (left) or 0.5 (right): * for 10%, ** for 5%, *** for 1%.

4.3 Fitting the Rabin and Vayanos Model

Given the results from the previous subsection, the analysis now proceeds along lines that differentiate between different types of subjects, especially between 'gf' and 'hh' types. In this subsection, the Rabin and Vayanos (2010) model presented in subsection 2.3 is fitted to the data for all sequences and responses from 'gf' type subjects, while controling for sequence length and round.[16] The fitted model is then given by

[16]As shown in subsection 4.6, for 'hh' type subjects, this is not the correct functional form, so an appropriate analysis for this type is delayed until then.

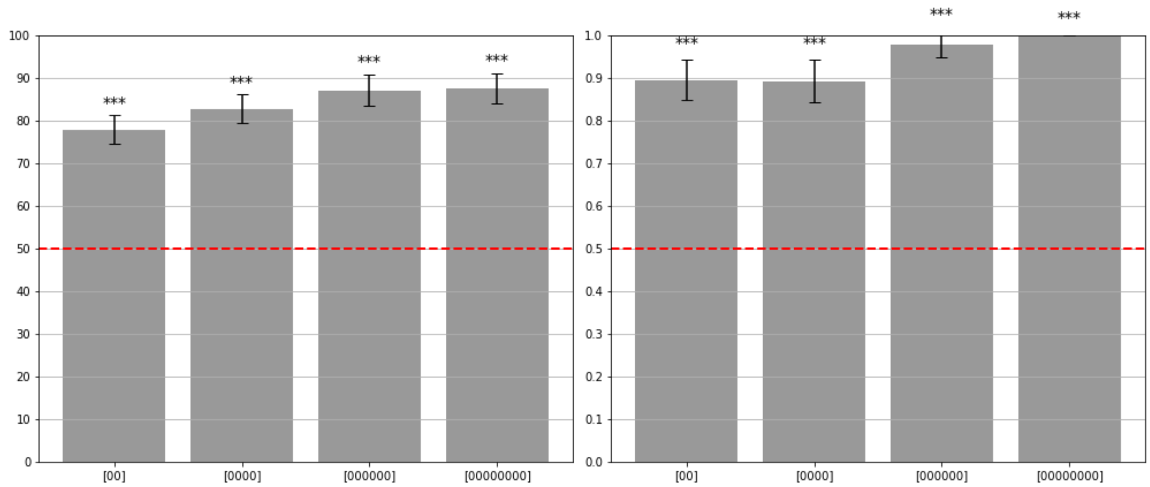


Figure 6: Mean probability (left) and choice frequency (right) responses for the outcome that continues the full run. 'Hot hands' subjects. Error bars are 95% confidence intervals. Stars record level of significance at which the displayed values are different from 50 (left) or 0.5 (right): * for 10%, ** for 5%, *** for 1%.

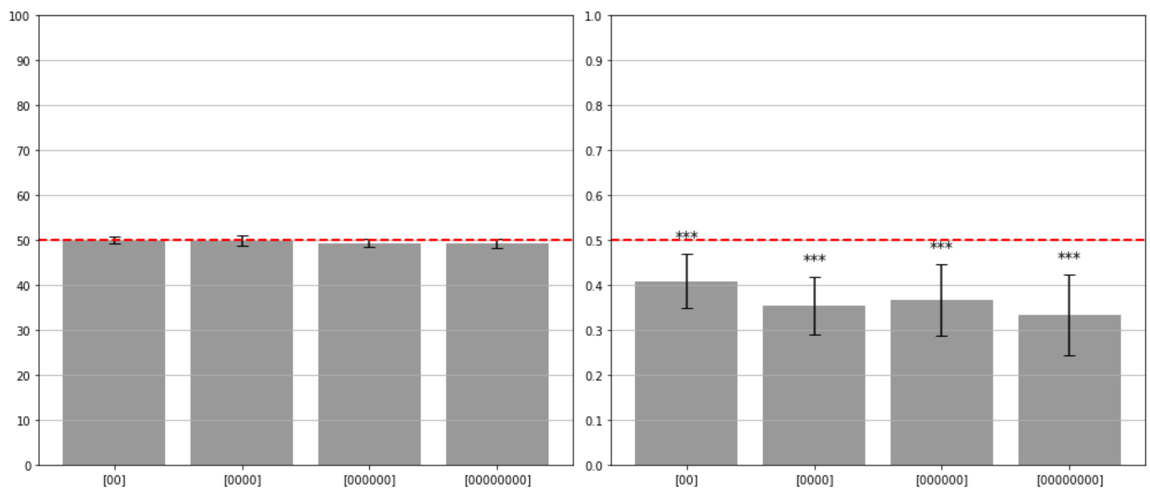


Figure 7: Mean probability (left) and choice frequency (right) responses for the outcome that continues the full run. 'Rational' subjects. Error bars are 95% confidence intervals. Stars record level of significance at which the displayed values are different from 50 (left) or 0.5 (right): * for 10%, ** for 5%, *** for 1%.

$$q_{0j} = \sum_{i=1}^{8} \left(\alpha \delta^{i} pos_{ij} \right) + \beta_{length} length_j + \beta_{round} round_j + \varepsilon_j \tag{1}$$

where q_0 is the reported probability of 0 being the next outcome, transformed to

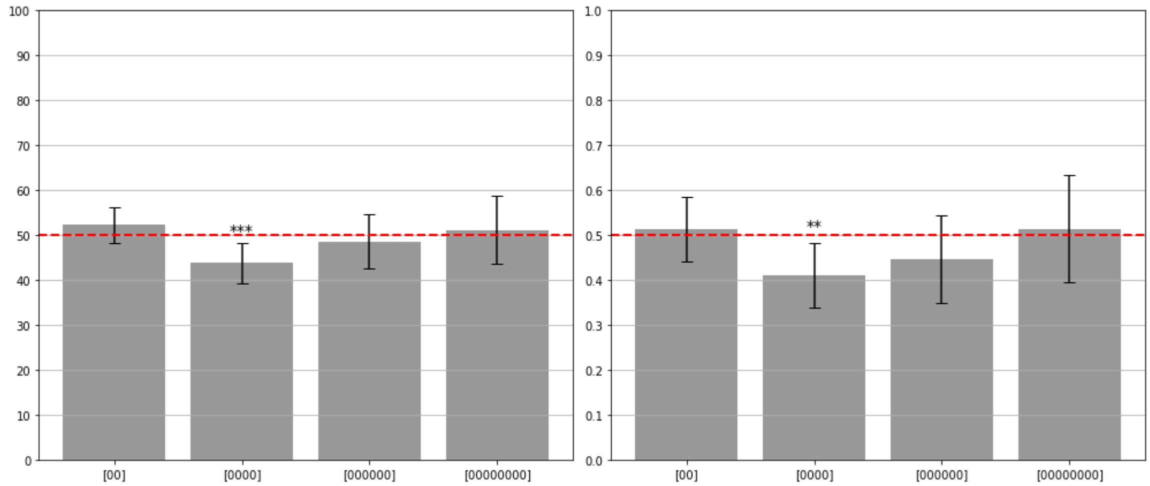


Figure 8: Mean probability (left) and choice frequency (right) responses for the outcome that continues the full run. 'Both' subjects. Error bars are 95% confidence intervals. Stars record level of significance at which the displayed values are different from 50 (left) or 0.5 (right): * for 10%, ** for 5%, *** for 1%.

$[-1.1]$, $length \in \{0, 2, 4, 6, 8\}$, $round \in \{1, 2, ..., 53, 54\}$, and $pos_1, pos_2, ..., pos_8$ are tertiary dummies encoding the value of the outcomes in each position according to

$$pos_i = \begin{cases} 1, \text{ if } = 1, \\ 0, \text{ if } = \emptyset, \\ -1, \text{ if } = 0. \end{cases}$$

Table 3 reports the estimated parameters from a nonlinear least squares estimation. Both the reversal parameter $\alpha = 0.224$ and the recency parameter $\delta = 0.795$ are of a significant size, suggesting that both effects play an important role in subjects' responses. This is only an analysis at an aggregate level, and a closer inspection in subsection 4.5 will indeed confirm this and provide additional details.

4.4 Empirical Entropies

This subsection follows the theoretical results from subsection 2.2 to investigate the effect of empirical entropies on reported beliefs. In order to do so, a regression is

Parameter	Estimate
α	0.224324***
	(0.015481)
δ	0.795059***
	(0.021439)
β_{length}	-0.000587
	(0.001557)
β_{round}	-0.000026
	(0.000281)
N	4590

Table 3: Results for nonlinear least squares estimation of the Rabin and Vayanos (2010) model for 'gf' type subjects according to equation 1. Standard errors clustered at the subject level. * significant at 10% level, ** significant at 5% level, *** significant at 1% level.

run in order to obtain linear coefficients for the delta empirical entropies $\Delta_0 H_k$, $k = 0, 1, 2$, on reported beliefs of the next outcome being 0, transformed to $[-1, 1]$, while controlling for sequence length and round. All rounds are used in the regression. This is done separately for subjects of types 'gf' and 'hh'. The regression equation is given by

$$q_{0j} = \alpha + \beta_0 \Delta_0 H_{0j} + \beta_1 \Delta_0 H_{1j} + \beta_2 \Delta_0 H_{2j} + \beta_3 length_j + \beta_4 round_j + \varepsilon_j$$

Remember that $\Delta_0 H_k(s^n) = H_k(s^n 0) - H_k(s^n 1)$ and that, for the gambler's fallacy, it is expected that the outcome that makes the sequence more complex are more likely. Since the dependent variable is the probability of 0, for 'gf' type subjects, the expected signs of the coefficients are $\beta_0 > 0$, $\beta_1 > 0$ and $\beta_2 > 0$. If 'hh' type subjects are the mirror of 'gf' subjects, one would expect that $\beta_0 < 0$, $\beta_1 < 0$ and $\beta_2 < 0$.

Table 4 reports the results of these regressions. For 'gf' type subjects, the effects of empirical entropies are all highly significant, and their effect goes in the expected direction. The larger the increase in complexity that outcome 0 makes when compared to outcome 1, the more likely 'gf' subjects think outcome 0 is to come next. Another thing to observe from the results is that $\beta_0 > \beta_1 > \beta_2$, meaning that lower order empirical entropies have a larger effect on beliefs.[17] This is to be expected if one considers that higher order entropies capture more complicated patterns, which subjects might find harder to notice or reason about.

For ‘hh’ type subjects, the coefficients also have the expected sign. It’s still the case that the coefficient for order zero is the largest one, but the coefficient for order one is not as significant and is the smallest one. This continues a broader trend of ‘hh’ type subjects not being perfect mirrors of ‘gf’ type subjects.

	q_0	
	(gf)	(hh)
const	-0.0259	-0.0180
	(0.016)	*(0.031)*
$\Delta_0 H_0$	0.5856***	-0.7366***
	(0.037)	*(0.059)*
$\Delta_0 H_1$	0.2621***	-0.1186*
	(0.036)	*(0.062)*
$\Delta_0 H_2$	0.1254***	-0.2536***
	(0.044)	*(0.076)*
length	0.0026	-0.0011
	(0.003)	*(0.004)*
rounds	0.0002	0.0009
	(0.000)	*(0.001)*
R^2	0.308	0.268
N	4590	3132

Table 4: Regression results for ‘gf’ and ‘hh’ type subjects. q_0, the dependent variable, is the reported probability of 0 as the next outcome. OLS regression with standard errors clustered at the subject level. * significant at 10% level, ** significant at 5% level, *** significant at 1% level.

4.5 Closer Look: Individual Sequences

The previous two subsections have shown that at an aggregate level, both theories presented in section 2 perform well when explaining the gambler’s fallacy effects found

[17] The null hypothesis that $\beta_0 \leq \beta_1$ is rejected at the 1% level, and the null hypothesis that $\beta_1 \leq \beta_2$ is rejected at the 5% level.

in subjects‘ behavior. The Rabin and Vayanos (2010) model results in significant coefficients that confirm the reserval and recency effects. While empirical entropies have been shown to have a highly significant effect on beliefs, with outcomes that increase sequence complexity being seen as more likely. This subsection, therefore, moves the analysis from an aggregate level down to the level of individual sequences in order to take a closer look at subject behavior and allow the differentiation between the two theories.

Appendix A provides the full look at all sequences of length 2, 4 and 6. This subsection focuses on the cases that provide the sharpest differentiation between the two theories. In particular, it presents the case of the *alternating sequences* (0101... and 1010...), for which subjects predict a continuation of the pattern rather than its break. This violates the predictions made by the entropy model, while still closely following the predictions of the Rabin and Vayanos (2010) model. It also presents the case of the sequences for which both outcomes affect complexity equally, and so no prediction comes from the entropy model. Nonetheless, subjects still, for the most part, deviate from correct beliefs. The most important effect driving these results is the recency effect created by the discount factor δ, and so this effect is isolated for the cases of sequences with just one different outcome, but in different positions.

The figures in this subsection are similar to the ones shown in subsection 4.1. Probability responses are presented on the left and choice responses on the right. Except when noted, these responses are for the outcome that results in lower entropy. The red dots shown on the left panels mark the point-prediction of the Rabin and Vayanos model for the sequence when using the parameters estimated in subsection 4.3 ($(\alpha, \delta) = (0.224, 0.795)$). All results are for ‘gf’ type subjects only.

4.5.1 Length 4: $\Delta H_0 = 0$, $\Delta H_1 > 0$, $\Delta H_2 = 0$

Figure 9 shows responses for low entropy outcome for length 4 sequences that have $\Delta H_0 = 0$, $\Delta H_1 > 0$, $\Delta H_2 = 0$. It is among these sequences that some of the most significant violations of the entropy theory happen. For sequences 0101, 1010, 0110 and 1001, subjects believe that the low entropy outcome is more likely to come next. While these responses violate the entropy prediction, they are in line with the prediction made by the Rabin and Vayanos model. For example, for the alternating sequences, 0101 and 1010, because of the recency effect, one of the outcomes is further to the front in all instances and so gets overweighted relative to the other one, leading subjects to believe it less likely to come next.

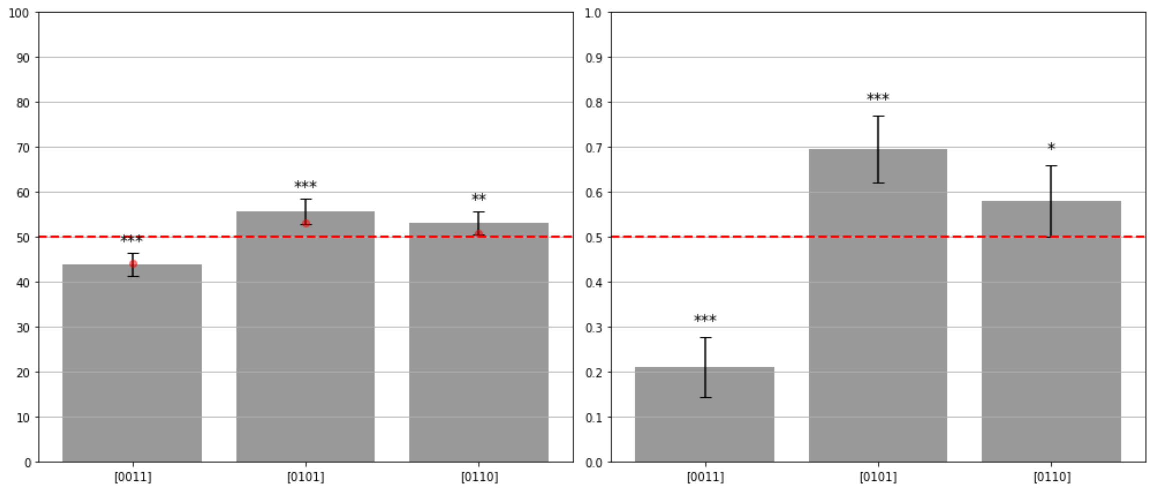


Figure 9: Mean probability (left) and choice frequency (right) responses for lower entropy outcome for length 4 sequences with $\Delta H_0 = 0$, $\Delta H_1 > 0$, $\Delta H_2 = 0$. 'gf' subjects. Error bars are 95% confidence intervals. Stars record level of significance at which the displayed values are different from 50 (left) or 0.5 (right): * for 10%, ** for 5%, *** for 1%.

4.5.2 Length 6: $\Delta H_0 = 0$, $\Delta H_1 > 0$, $\Delta H_2 > 0$

Figure 10 shows responses for low entropy outcome for length 6 sequences that have $\Delta H_0 = 0$, $\Delta H_1 > 0$, $\Delta H_2 > 0$. Although at a lower significance level, the alternating sequence here further confirms the same effect as seen in the previous figure. Subjects responses go in the opposite direction to that predicted by the entropy model, while being in line with the prediction of the Rabin and Vayanos model.

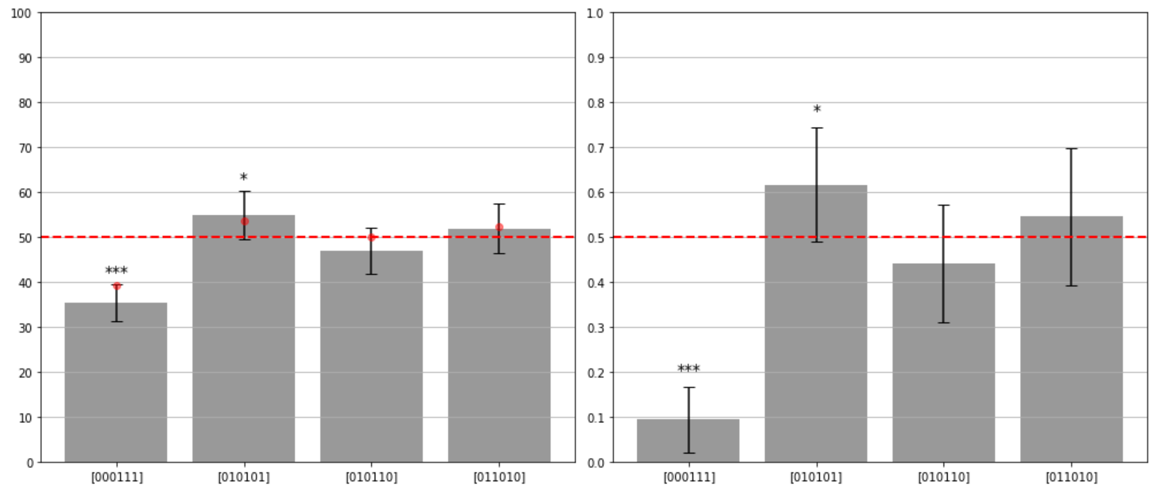


Figure 10: Mean probability (left) and choice frequency (right) responses for lower entropy outcome for length 6 sequences with $\Delta H_0 = 0$, $\Delta H_1 > 0$, $\Delta H_2 > 0$. 'gf' subjects. Error bars are 95% confidence intervals. Stars record level of significance at which the displayed values are different from 50 (left) or 0.5 (right): * for 10%, ** for 5%, *** for 1%.

4.5.3 Length 6: $\Delta H_0 = \Delta H_1 = \Delta H_2 = 0$

Figure 11 shows responses for outcome 0 for length 6 sequences that have $\Delta H_0 = \Delta H_1 = \Delta H_2 = 0$. That is, these are sequences for which there is no 'low entropy' outcome, as both outcomes 0 and 1 result in a sequence with same empirical entropy of orders 0, 1 and 2. These can be thought of sequences that are already highly complex, and there's little pattern to be matched or reinforced by either outcome. Nevertheless, it's shown that for most of these sequences, subjects still deviate significantly from the correct belief. Their deviation closely follows the prediction of Rabin and Vayanos. It's worth noting that this happens with sequences that have the same outcome repeated twice at the end of the sequence, further reinforcing the importance of the recency effect.

4.5.4 Lengths 4 and 6: recency effects

Figures 12 and 13 illustrate in more detail the recency effect for sequences of length 4 and 6, respectively. They show in order sequences that have one (or no) outcome

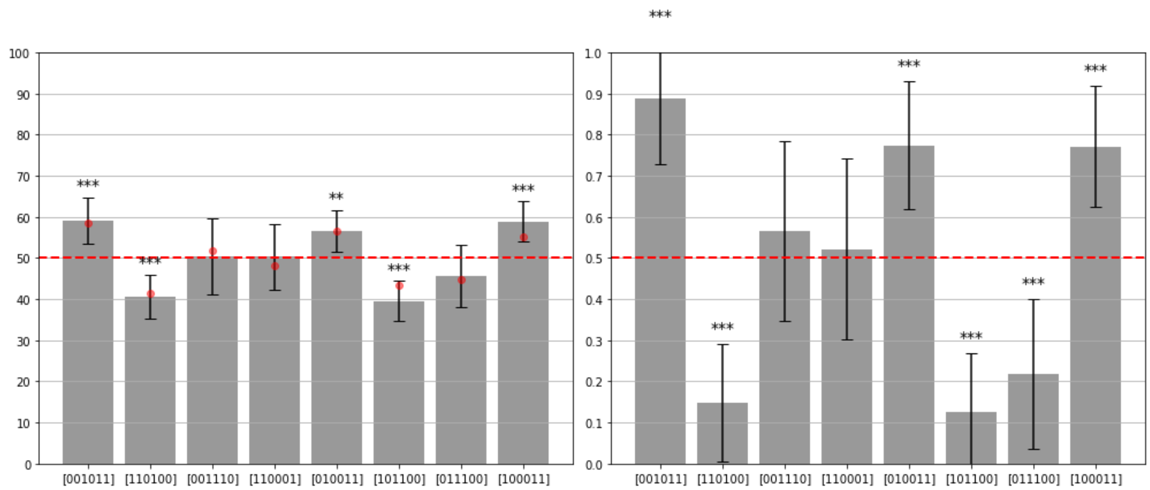


Figure 11: Mean probability (left) and choice frequency (right) responses for outcome 0 for length 6 sequences with $\Delta H_0 = 0$, $\Delta H_1 = 0$, $\Delta H_2 = 0$. 'gf' subjects. Error bars are 95% confidence intervals. Stars record level of significance at which the displayed values are different from 50 (left) or 0.5 (right): * for 10%, ** for 5%, *** for 1%.

different than the rest, but differing by the position in which it appears. These figures show a monotonicity that is created by the recency effect. The closer the different outcome is to the front of the sequence, the larger it's impact in countering the effect of the rest of the outcomes on the belief. For example, subjects believe 0 is less likely to be next when facing the sequence 000001 than the sequence 100000. This cannot be explained by empirical entropy, but is in line with the Rabin and Vayanos model.

4.6 Generalized Rabin and Vayanos

The previous subsection has shown that the Rabin and Vayanos (2010) model has a superior performance in matching the actual behavior of 'gf' type subjects. This subsection explores whether it is possible to increase the performance of the model in a significant way by increasing the number of parameters, creating a generalized version of their model.

The generalized version of the model takes the form

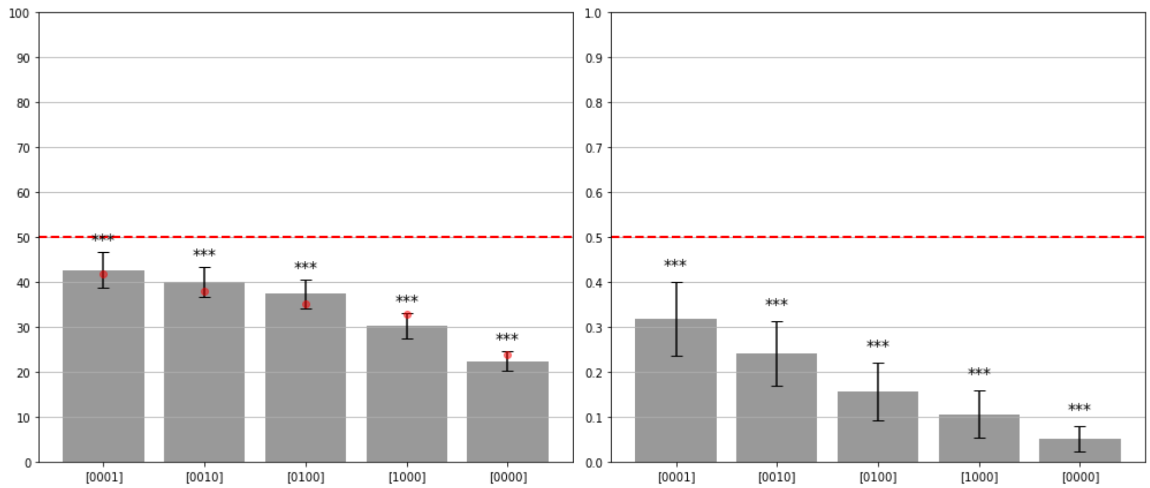


Figure 12: Mean probability (left) and choice frequency (right) responses for lower entropy outcome, showing recency effect for length 4 sequences. 'gf' subjects. Error bars are 95% confidence intervals. Stars record level of significance at which the displayed values are different from 50 (left) or 0.5 (right): * for 10%, ** for 5%, *** for 1%.

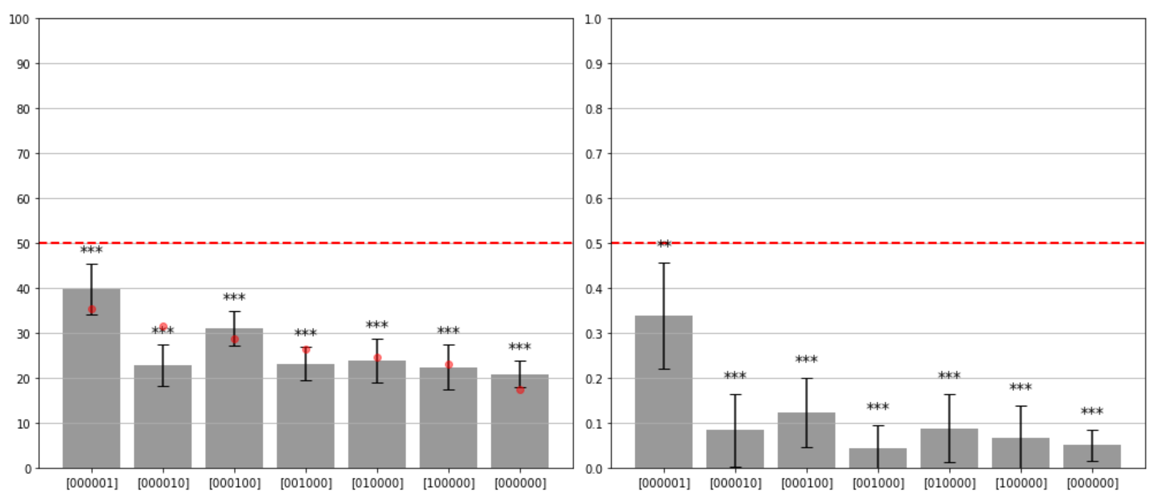


Figure 13: Mean probability (left) and choice frequency (right) responses for lower entropy outcome, showing recency effect for length 6 sequences. 'gf' subjects. Error bars are 95% confidence intervals. Stars record level of significance at which the displayed values are different from 50 (left) or 0.5 (right): * for 10%, ** for 5%, *** for 1%.

$$q_{0j} = \sum_{i=1}^{8} (\beta_i pos_{ij}) + \beta_{length} length_j + \beta_{round} round_j + \varepsilon_j. \qquad (2)$$

Before, the model had only two parameters, α and δ, and it was the case that they implied $\beta_i = \alpha\delta^i$. These implied parameters can be be compared to the ones that are estimated separately in the generalized model. Table 5 reports the results of this estimation for subject types 'gf' and 'hh'. The position parameters are all highly significant and with the expected signs.

Figure 14 shows the comparison of the estimated values of the general parameters and the parameters implied by the previously estimated α and δ for 'gf' type subjects. The conclusion is that the generalization makes little difference. The free estimation of the 8 parameters result in numbers that are very close to the numbers that were already implied by α and δ. This conclusion is further reinforced by a likelihood ratio test between the two models, which returns a p-value of 0.169, failing to reject even at the 10% significance level the null hypothesis that the 8 parameter model provides no additional explanatory power compared to the 2 parameter model.

Figure 15 plots the estimated parameters for 'gf' and 'hh' subjects. It highlights the difference in behavior for 'hh' subjects. While for 'gf' subjects the parameters show the recency bias implied by the discount factor, with more recent results being weighted more heavily, for 'hh' subjects, it's actually the outcomes in the middle of the sequence that are more heavily weighted.

5 Discussion and Conclusion

In this paper I have compared the performance of two theoretical models of the gambler's fallacy: a formalized version of the representativeness heuristic using information theory, and the recency-weighted reversal model of Rabin and Vayanos (2010). Although both models perform well when taking an aggregate look at the data, more detailed analysis at the level of individual sequence reveals that the latter has a superior performance.

The main component of this superior performance comes from the clear recency

	q_0	
	(gf)	(hh)
const	-0.0261	-0.0035
	(0.016)	*(0.029)*
pos_1	0.1849***	-0.0455**
	(0.011)	*(0.021)*
pos_2	0.1344***	-0.1023***
	(0.010)	*(0.015)*
pos_3	0.1099***	-0.1377***
	(0.008)	*(0.012)*
pos_4	0.0800***	-0.1514***
	(0.010)	*(0.014)*
pos_5	0.0796***	-0.1165***
	(0.009)	*(0.014)*
pos_6	0.0646***	-0.1298***
	(0.008)	*(0.011)*
pos_7	0.0477***	-0.1035***
	(0.014)	*(0.022)*
pos_8	0.0425***	-0.0528***
	(0.012)	*(0.017)*
rounds	0.0003	0.0010
	(0.000)	*(0.001)*
length	0.0024	-0.0032
	(0.003)	*(0.004)*
R^2	0.425	0.276
N	4590	3132

Table 5: Regression results for 'gf' and 'hh' type subjects of the generalized Rabin and Vayanos model. q_0, the dependent variable, is the reported probability of 0 as the next outcome. OLS regression with standard errors clustered at the subject level. * significant at 10% level, ** significant at 5% level, *** significant at 1% level.

bias that subjects exhibit in their behavior. Subjects' beliefs were more responsive to the outcomes appearing more to the front of the sequences, whereas those further back had a lesser effect. Previous experiments, which did not collect belief responses for general sequences, could not definitively confirm this bias in the context of the gambler's fallacy. This behavior adds to the long standing literature on recency bias going back to Sternberg (1966). Afrouzi et al. (2023) provide, along with their own model and experimental results, an overview of the literature. Both in their model

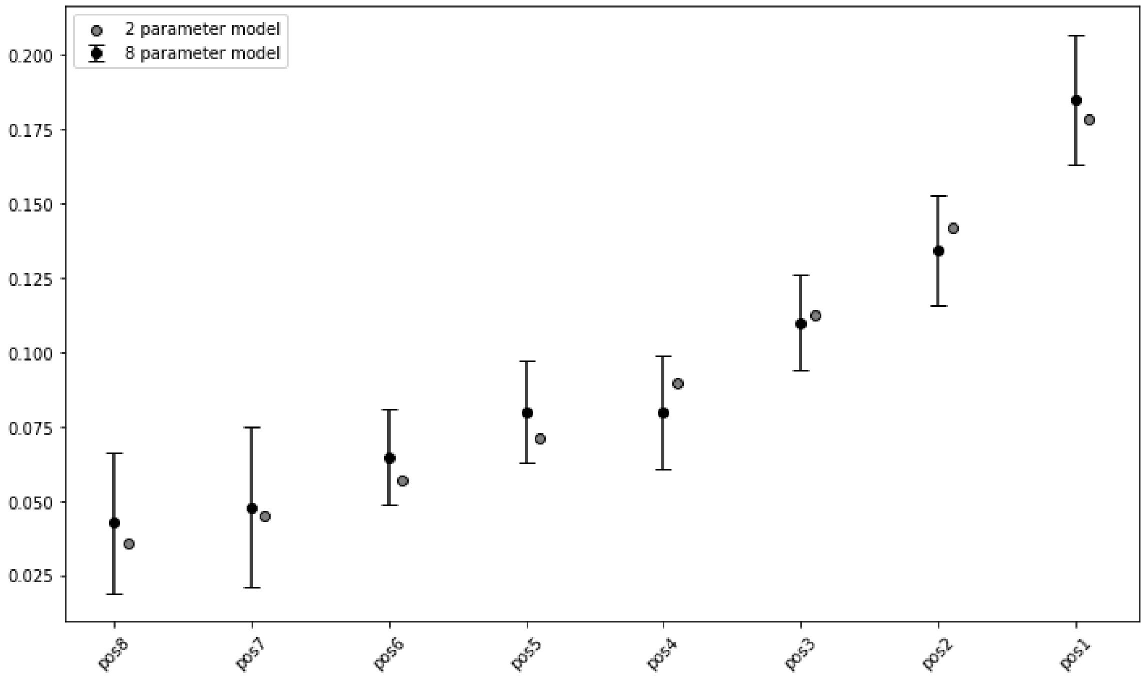


Figure 14: Comparison of the estimated 8 parameters in the generalized Rabin and Vayanos model and the implied 2 parameters original version. Error bars are 95% confidence intervals. 'gf' type subjects.

and in more recent work they survey,[18] the recency bias is modeled as resulting from costly cognitive processing of information in working memory formation. This extends to cases in which no recall is actually required, as in this paper's experiment, since there's an important component of attention control to working memory (Unsworth and Spillers 2010). Future work on the gambler's fallacy will have to engage with this literature on the recency bias, potentially explicitly taking into account its foundations in costly cognition.

Natural extensions of the current work would be to expand the space of either possible outcomes or responses, or the underlying data generating process more generally. One such extension which is planned for this project is to run a very similar experiment with a fair coin, but ask subjects for the probabilities of the next two outcomes (that is, a belief $q \in \Delta(\{00, 01, 10, 11\})$). It's an open question which of the results from the current experiment will hold in this new version. In particular,

[18]See also their online appendix.

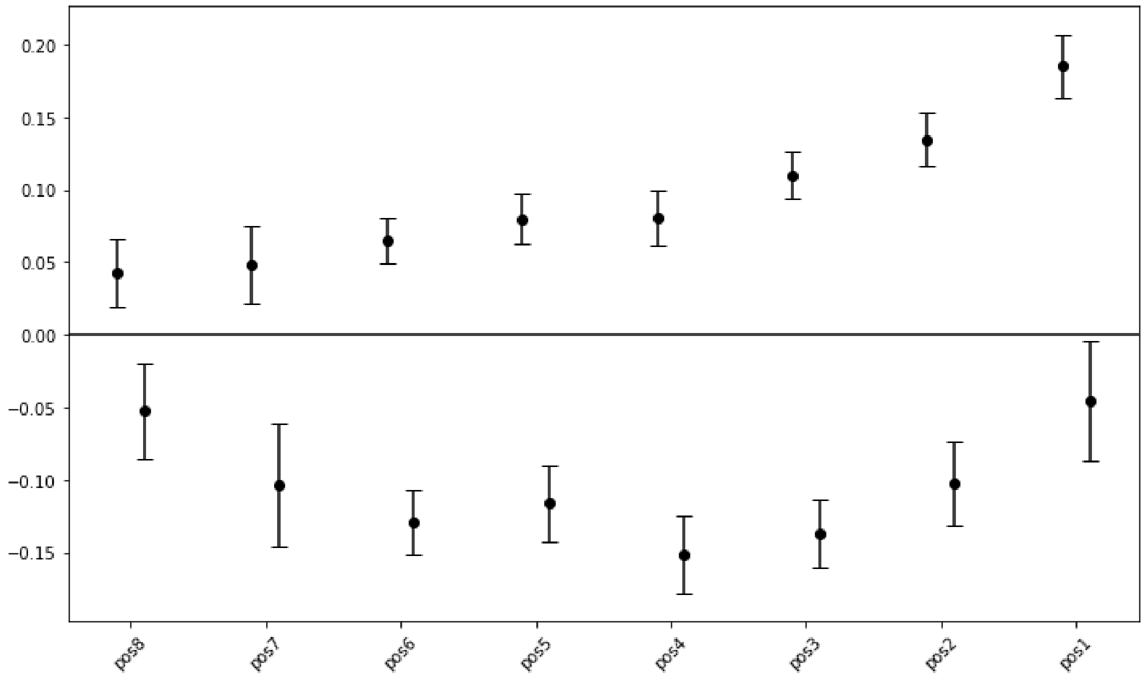


Figure 15: Comparison of estimated parameters of the generalized Rabin and Vayanos model for 'gf' type subjects (top) and 'hh' type subjects (bottom). Error bars are 95% confidence intervals.

what will subjects say is the most likely of the four outcomes when facing the alternating sequences *head-tail-head-tail...*. If the Rabin and Vayanos model extends to this setup, subjects should still think *head-tail* is most likely next. It might be, however, that in this case it becomes more evident that they are continuing the first order pattern of the sequence and therefore judge another outcome, which increases complexity, as more likely to happen. Other extensions would be to use a biased, instead of a fair coin, or a process with more than two outcomes to go beyond the binary case.

Finally, the complexity measure of higher order empirical entropy which has received attention here might find other uses as a more general measure of information complexity. It might find applications in measuring costly information processing and its relations to information preferences.

References

Afrouzi, Hassan, Spencer Y Kwon, Augustin Landier, Yueran Ma, and David Thesmar, "Overreaction in Expectations: Evidence and Theory," *The Quarterly Journal of Economics*, March 2023, *138* (3), 1713–1764.

Bar-Hillel, Maya, "What features make samples seem representative?," *Journal of Experimental Psychology: Human Perception and Performance*, August 1980, *6* (3), 578–589.

_ , "Representativeness and fallacies of probability judgment," *Acta Psychologica*, March 1984, *55* (2), 91–107.

Benjamin, Daniel, Don Moore, and Matthew Rabin, *Biased Beliefs About Random Samples: Evidence from Two Integrated Experiments* October 2017.

Dale, Andrew I. and Pierre-Simon Laplace, *Philosophical Essay on Probabilities*, Springer New York, 1825/1995.

Danz, David, Lise Vesterlund, and Alistair J. Wilson, "Belief Elicitation and Behavioral Incentive Compatibility," *American Economic Review*, September 2022, *112* (9), 2851–2883.

de Leeuw, J. R., "jsPsych: A JavaScript library for creating behavioral experiments in a web browser," *Behavior Research Methods*, 2015, *47* (2), 1–12.

Ferragina, Paolo and Giovanni Manzini, "Indexing compressed text," *Journal of the ACM*, July 2005, *52* (4), 552–581.

Gagie, Travis, "Large alphabets and incompressibility," *Information Processing Letters*, September 2006, *99* (6), 246–251.

Grether, David M., "Bayes Rule as a Descriptive Model: The Representativeness Heuristic," *The Quarterly Journal of Economics*, November 1980, *95* (3), 537.

Hossain, T. and R. Okui, "The Binarized Scoring Rule," *The Review of Economic Studies*, January 2013, *80* (3), 984–1001.

Kahneman, Daniel and Amos Tversky, "Subjective probability: A judgment of representativeness," *Cognitive Psychology*, July 1972, *3* (3), 430–454.

Manzini, Giovanni, "An analysis of the Burrows—Wheeler transform," *Journal of the ACM*, May 2001, *48* (3), 407–430.

Oskarsson, An T., Leaf Van Boven, Gary H. McClelland, and Reid Hastie, "What's next? Judging sequences of binary events.," *Psychological Bulletin*, 2009, *135* (2), 262–285.

Rabin, Matthew, "Inference by Believers in the Law of Small Numbers," *The Quarterly Journal of Economics*, August 2002, *117* (3), 775–816.

_ **and Dimitri Vayanos**, "The Gambler's and Hot-Hand Fallacies: Theory and Applications," *Review of Economic Studies*, April 2010, *77* (2), 730–778.

Shannon, C. E., "A Mathematical Theory of Communication," *Bell System Technical Journal*, July 1948, *27* (3), 379–423.

Skinner, B. F., "The processes involved in the repeated guessing of alternatives.," *Journal of Experimental Psychology*, June 1942, *30* (6), 495–503.

Sternberg, Saul, "High-Speed Scanning in Human Memory," *Science*, August 1966, *153* (3736), 652–654.

Tenenbaum, Joshua B. and Thomas Griffiths, "The rational basis of representativeness," in J. Moore and K. Stenning, eds., *Proceedings of the 23rd Annual Conference of the Cognitive Science Society*, Erlbaum Mahwah, NJ 2001, pp. 1036–1041.

Tversky, Amos and Daniel Kahneman, "Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty.," *Science*, September 1974, *185* (4157), 1124–1131.

_ **and** _ , "Judgments of and by representativeness," in Daniel Kahneman, Paul Slovic, and Amos Tversky, eds., *Judgment under Uncertainty*, Cambridge University Press, 1982, pp. 84–98.

Unsworth, Nash and Gregory J. Spillers, "Working memory capacity: Attention control, secondary memory, or both? A direct test of the dual-component model," *Journal of Memory and Language*, May 2010, *62* (4), 392–406.

A Individual Sequences

This appendix provides the complete look at individual sequences of lengths 2, 4 and 6.

Figures are similar to the ones shown in subsection 4.1 of the main text. Probability responses are presented on the left and choice responses on the right. Except when noted, these responses are for the outcome that results in lower entropy. The red dots shown on the left panels mark the point-prediction of the Rabin and Vayanos model for the sequence when using the parameters estimated in subsection 4.3 ($(\alpha, \delta) = (0.224, 0.795)$).

Sequences are organized according to their length and also on the orders for which empirical entropy matters. For example, for some sequences, outcomes 0 and 1 only make a difference at the level of empirical entropy of order 1. That is, they are sequences s such that $\Delta H_1(s) > 0$ but $\Delta H_0(s) = 0$ and $\Delta H_2(s) = 0$. All results are for 'gf' type subjects only.

A.1 Length 0

Throughout the experiment, subjects encountered twice sequences of length 0. That is, the screen had no outcomes shown, and subjects were asked to bet on the first outcome being a head or a tail, as well as give the probabilities that the first outcome is head or tail. Figure A.1 shows responses when subjects encountered this length 0 sequences for the first time, for the second time, and aggregating across both. Responses are for outcome 0. The slight bias towards 1 (head) probably reflects the fact that in the task screen, this outcome was lower and therefore closer to the confirmation button at the bottom. (See figure 1.) Importantly, the second time they encountered such sequences, subjects were more likely to be careful in giving the correct 50-50 answer.

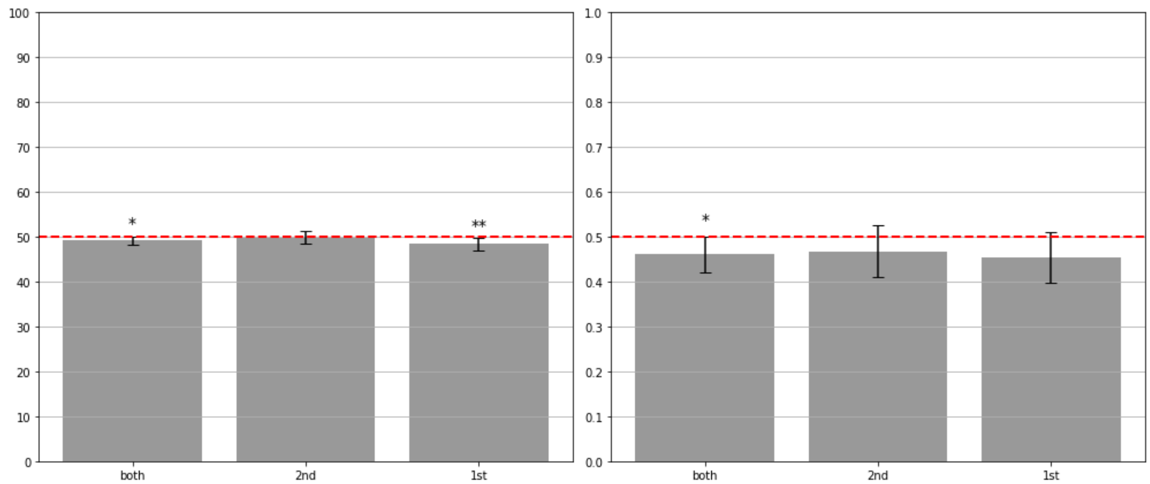


Figure A.1: Mean probability (left) and choice frequency (right) responses for 0 when encountering sequences of length 0 for the first and second times. All subjects. Error bars are 95% confidence intervals. Stars record level of significance at which the displayed values are different from 50 (left) or 0.5 (right): * for 10%, ** for 5%, *** for 1%.

A.2 Length 2

Figure A.2 shows responses for outcome 0 for the length 2 sequences. Although not significantly different from 50%, the mean probability responses for the sequences 10 and 01 already serve as a preview from the effect that recency has on subjects' responses.

A.3 Length 6: $\Delta H_0 > 0$, $\Delta H_1 > 0$, $\Delta H_2 > 0$

Figure A.3 shows responses for low entropy outcome for length 6 sequences that have $\Delta H_0 > 0$, $\Delta H_1 > 0$, $\Delta H_2 > 0$. For these sequences, both theories perform well. Subjects generally believe that low entropy outcomes are less likely to come next.

A.4 Length 6: $\Delta H_0 > 0$, $\Delta H_1 > 0$, $\Delta H_2 = 0$

Figure A.4 shows responses for low entropy outcome for length 6 sequences that have $\Delta H_0 > 0$, $\Delta H_1 > 0$, $\Delta H_2 = 0$. These sequences continue the theme from the previous

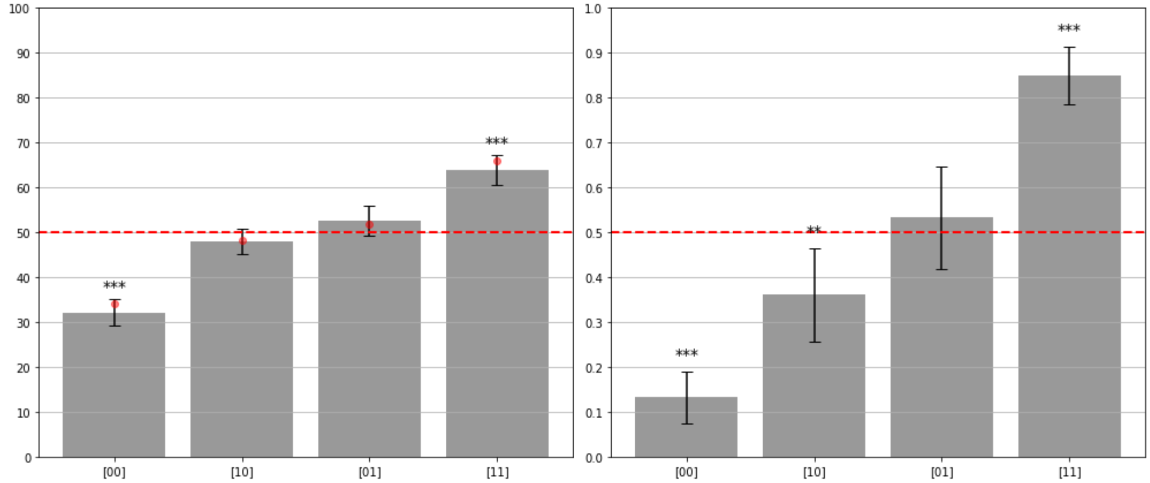


Figure A.2: Mean probability (left) and choice frequency (right) responses for outcome 0 for all length 2 sequences. ‘gf’ subjects. Error bars are 95% confidence intervals. Stars record level of significance at which the displayed values are different from 50 (left) or 0.5 (right): * for 10%, ** for 5%, *** for 1%.

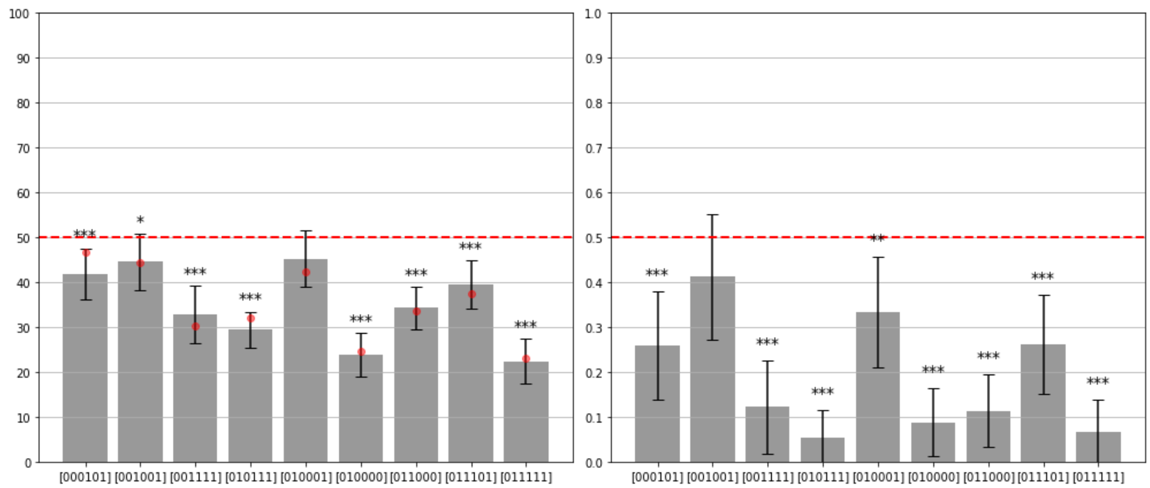


Figure A.3: Mean probability (left) and choice frequency (right) responses for lower entropy outcome for length 6 sequences with $\Delta H_0 > 0$, $\Delta H_1 > 0$, $\Delta H_2 > 0$. ‘gf’ subjects. Error bars are 95% confidence intervals. Stars record level of significance at which the displayed values are different from 50 (left) or 0.5 (right): * for 10%, ** for 5%, *** for 1%.

figure, with subjects thinking that low entropy outcomes are less likely.

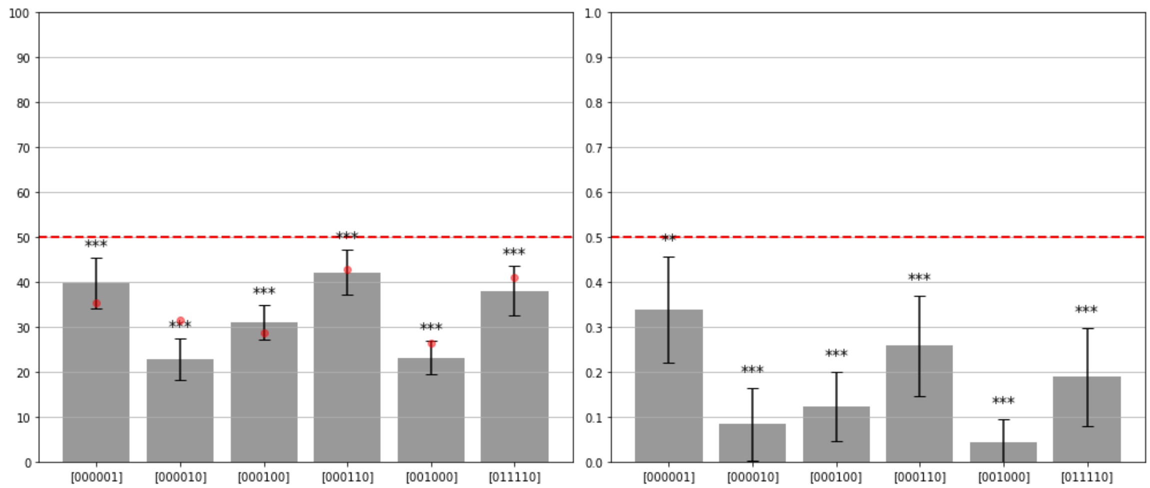


Figure A.4: Mean probability (left) and choice frequency (right) responses for lower entropy outcome for length 6 sequences with $\Delta H_0 > 0$, $\Delta H_1 > 0$, $\Delta H_2 = 0$. 'gf' subjects. Error bars are 95% confidence intervals. Stars record level of significance at which the displayed values are different from 50 (left) or 0.5 (right): * for 10%, ** for 5%, *** for 1%.

A.5 Length 4: $\Delta H_0 = 0$, $\Delta H_1 > 0$, $\Delta H_2 = 0$

Figure A.5 shows responses for low entropy outcome for length 4 sequences that have $\Delta H_0 = 0$, $\Delta H_1 > 0$, $\Delta H_2 = 0$. It is among these sequences that some of the most significant violations of the entropy theory happen. For sequences 0101, 1010, 0110 and 1001, subjects believe that the low entropy outcome is more likely to come next. While these responses violate the entropy prediction, they are in line with the prediction made by the Rabin and Vayanos model. For example, for the alternating sequences, 0101 and 1010, because of the recency effect, one of the outcomes is further to the front in all instances and so gets overweighted relative to the other one, leading subjects to believe it less likely to come next.

A.6 Length 6: $\Delta H_0 = 0$, $\Delta H_1 > 0$, $\Delta H_2 > 0$

Figure A.6 shows responses for low entropy outcome for length 6 sequences that have $\Delta H_0 = 0$, $\Delta H_1 > 0$, $\Delta H_2 > 0$. Although at a lower significance level, the alternating sequence here further confirms the same effect as seen in the previous figure. Subjects

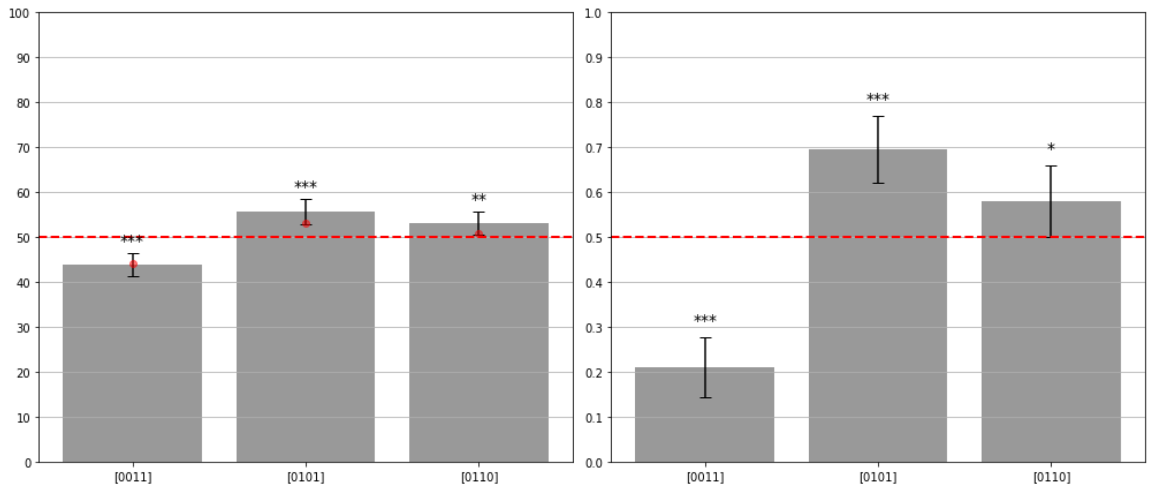


Figure A.5: Mean probability (left) and choice frequency (right) responses for lower entropy outcome for length 4 sequences with $\Delta H_0 = 0$, $\Delta H_1 > 0$, $\Delta H_2 = 0$. 'gf' subjects. Error bars are 95% confidence intervals. Stars record level of significance at which the displayed values are different from 50 (left) or 0.5 (right): * for 10%, ** for 5%, *** for 1%.

responses go in the opposite direction to that predicted by the entropy model, while being in line with the prediction of the Rabin and Vayanos model.

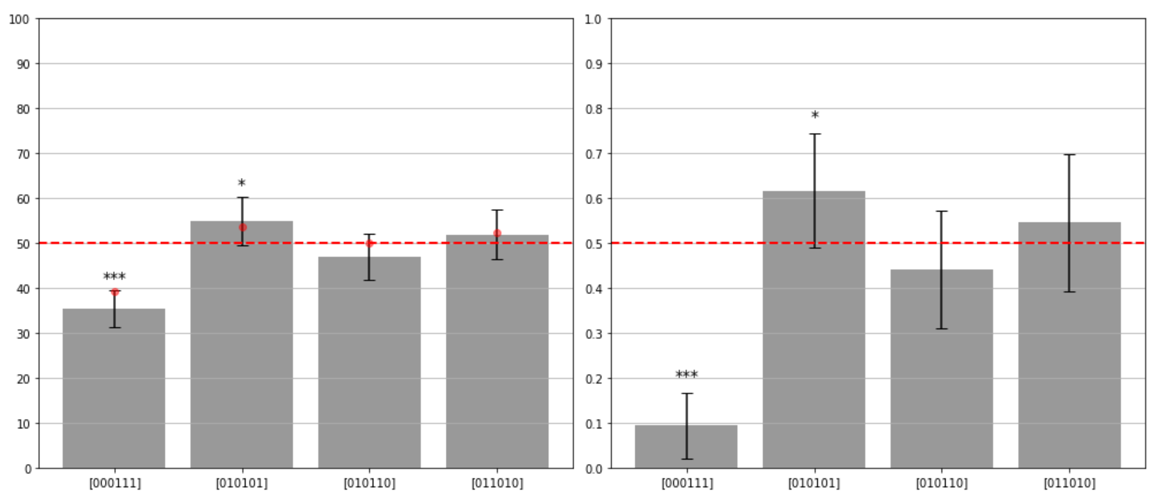


Figure A.6: Mean probability (left) and choice frequency (right) responses for lower entropy outcome for length 6 sequences with $\Delta H_0 = 0$, $\Delta H_1 > 0$, $\Delta H_2 > 0$. 'gf' subjects. Error bars are 95% confidence intervals. Stars record level of significance at which the displayed values are different from 50 (left) or 0.5 (right): * for 10%, ** for 5%, *** for 1%.

A.7 Length 6: $\Delta H_0 = 0$, $\Delta H_1 = 0$, $\Delta H_2 > 0$

Figure A.7 shows responses for low entropy outcome for length 6 sequences that have $\Delta H_0 = 0$, $\Delta H_1 = 0$, $\Delta H_2 > 0$. No significant effect is found for these sequences. Disaggregating the two symmetric sequences 011001 and 100110 makes no difference.

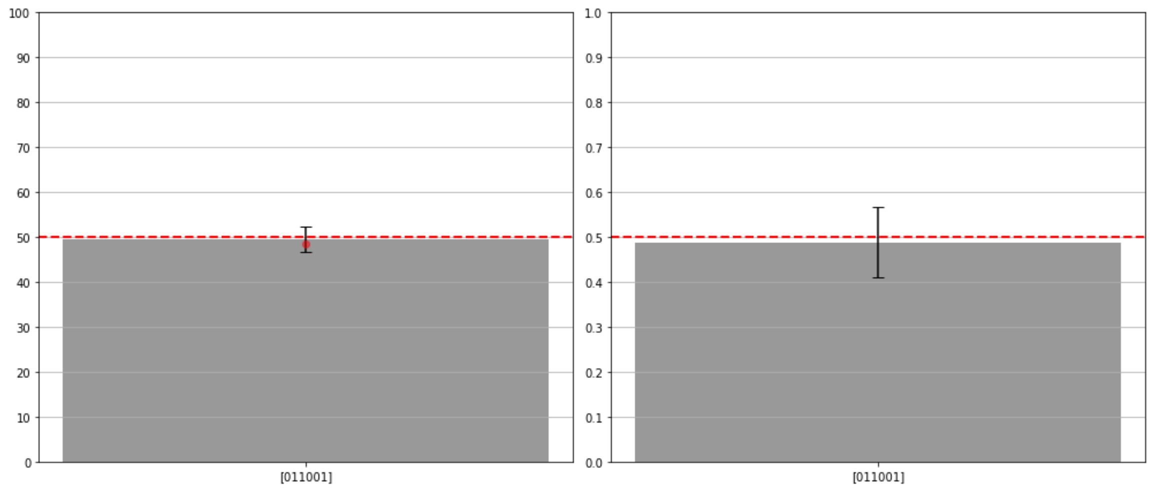


Figure A.7: Mean probability (left) and choice frequency (right) responses for lower entropy outcome for length 6 sequences with $\Delta H_0 = 0$, $\Delta H_1 = 0$, $\Delta H_2 > 0$. 'gf' subjects. Error bars are 95% confidence intervals. Stars record level of significance at which the displayed values are different from 50 (left) or 0.5 (right): * for 10%, ** for 5%, *** for 1%.

A.8 Length 6: ΔH_k with different signs

Figure A.8 shows responses for the outcome with low zero order entropy for length 6 sequences that have ΔH_k with different signs. That is, for these sequences, there is a tradeoff between increasing entropy at one order while decreasing it at another. Given the higher impact of lower order entropy, as shown by the estimated parameters in subsection 4.4, it might be expected that it is the outcome that increases zero order entropy that will be thought less likely to happen, which is indeed what is confirmed by these sequences.

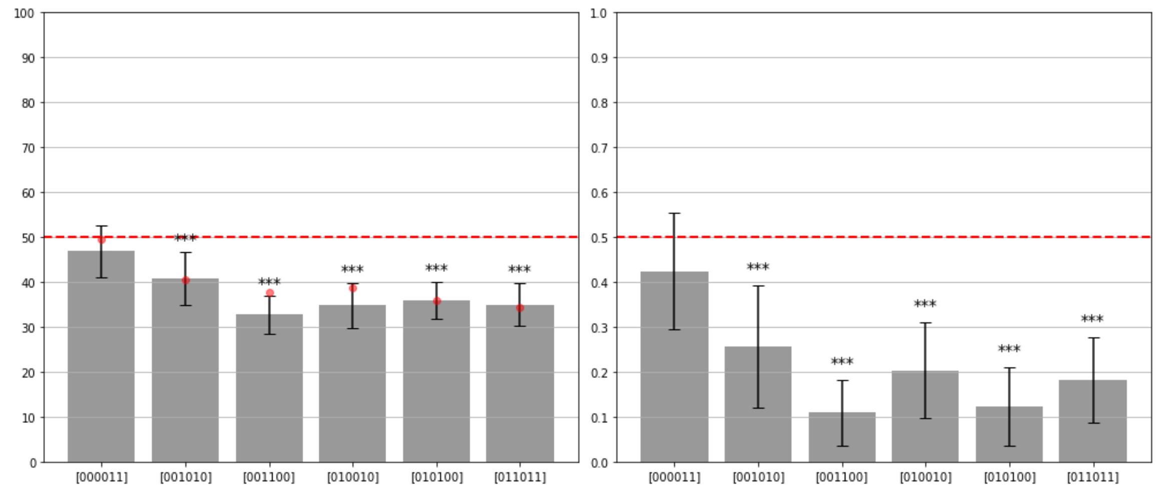


Figure A.8: Mean probability (left) and choice frequency (right) responses for lower zero order entropy outcome for length 6 sequences with ΔH_i having different signs. 'gf' subjects. Error bars are 95% confidence intervals. Stars record level of significance at which the displayed values are different from 50 (left) or 0.5 (right): * for 10%, ** for 5%, *** for 1%.

A.9 Length 6: $\Delta H_0 = \Delta H_1 = \Delta H_2 = 0$

Figure A.9 shows responses for outcome 0 for length 6 sequences that have $\Delta H_0 = \Delta H_1 = \Delta H_2 = 0$. That is, these are sequences for which there is no 'low entropy' outcome, as both outcomes 0 and 1 result in a sequence with same empirical entropy of orders 0, 1 and 2. These can be thought of sequences that are already highly complex, and there's little pattern to be matched or reinforced by either outcome. Nevertheless, it's shown that for most of these sequences, subjects still deviate significantly from the correct belief. Their deviation closely follows the prediction of Rabin and Vayanos. It's worth noting that this happens with sequences that have the same outcome repeated twice at the end of the sequence, further reinforcing the importance of the recency effect.

A.10 Lengths 4 and 6: recency effects

Figures A.10 and A.11 illustrate in more detail the recency effect for sequences of length 4 and 6, respectively. They show in order sequences that have one (or no)

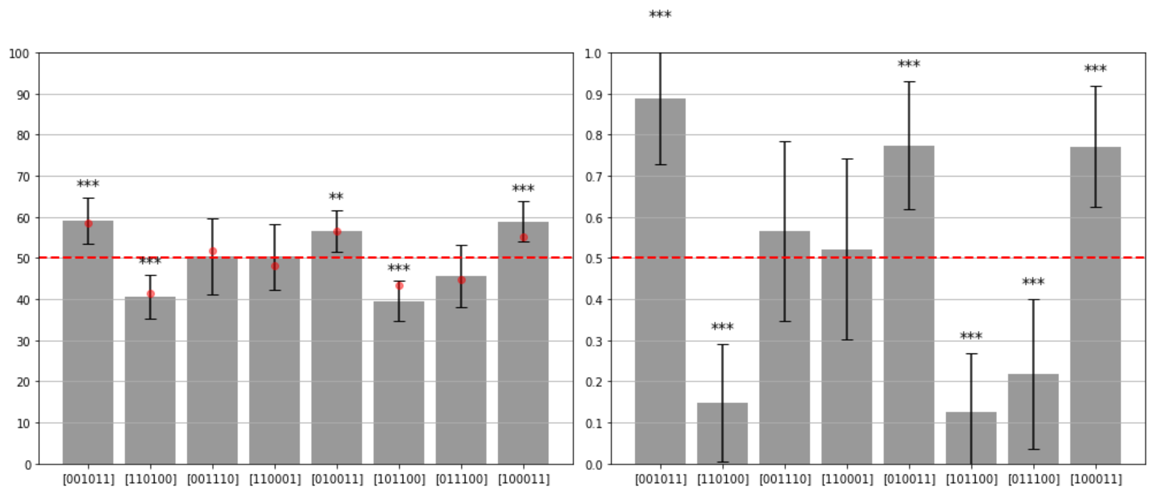


Figure A.9: Mean probability (left) and choice frequency (right) responses for outcome 0 for length 6 sequences with $\Delta H_0 = 0$, $\Delta H_1 = 0$, $\Delta H_2 = 0$. 'gf' subjects. Error bars are 95% confidence intervals. Stars record level of significance at which the displayed values are different from 50 (left) or 0.5 (right): * for 10%, ** for 5%, *** for 1%.

outcome different than the rest, but differing by the position in which it appears. These figures show a monotonicity that is created by the recency effect. The closer the different outcome is to the front of the sequence, the larger it's impact in countering the effect of the rest of the outcomes on the belief. For example, subjects believe 0 is less likely to be next when facing the sequence 000001 than the sequence 100000. This cannot be explained by empirical entropy, but is in line with the Rabin and Vayanos model.

B Aggregate Regressions

Given the subject heterogeneity explained in section 4.1 of the main text, the data analysis proceeded by treating 'gf' and 'hh' type subjects separately. This appendix shows results for all subjects, as well as 'rat' and 'both' subjects.

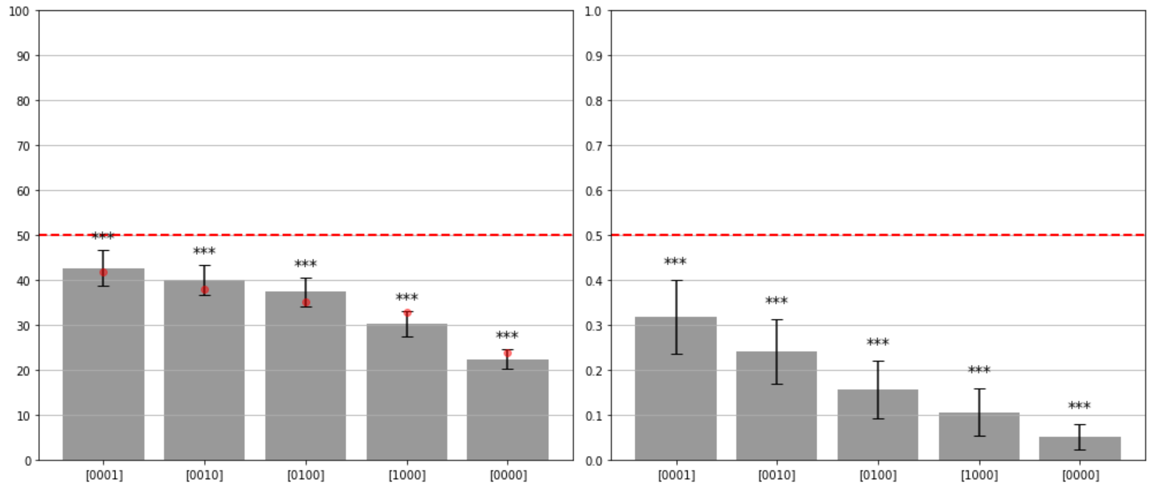


Figure A.10: Mean probability (left) and choice frequency (right) responses for lower entropy outcome, showing recency effect for length 4 sequences. 'gf' subjects. Error bars are 95% confidence intervals. Stars record level of significance at which the displayed values are different from 50 (left) or 0.5 (right): * for 10%, ** for 5%, *** for 1%.

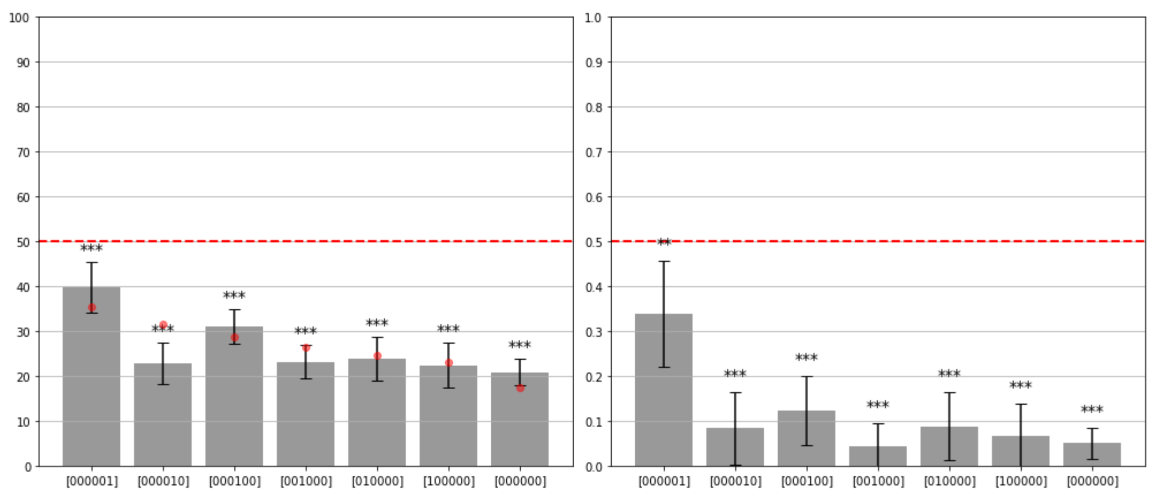


Figure A.11: Mean probability (left) and choice frequency (right) responses for lower entropy outcome, showing recency effect for length 6 sequences. 'gf' subjects. Error bars are 95% confidence intervals. Stars record level of significance at which the displayed values are different from 50 (left) or 0.5 (right): * for 10%, ** for 5%, *** for 1%.

B.1 Rabin and Vayanos model

Similarly to section 4.3 in the main text, tables B.1, B.2 and B.3 show fitted parameters for, respectively, all subjects, type ‘both’ subjects and type ‘rat’ subjects. As mentioned in the main text, this functional form is not adequate for ‘hh’ type subjects, and for these subjects the estimated parameters are essentially all 0. See table 5 and figure 15 in the main text for the appropriate analysis.

Parameter	Estimate
α	0.216846***
	(0.079945)
δ	0.297822**
	(0.121728)
β_{length}	0.000004
	(0.001054)
β_{rounds}	0.000115
	(0.000185)
N = 16038	

Table B.1: Results for nonlinear least squares estimation of the Rabin and Vayanos (2010) model for all subjects according to equation 1. Standard errors clustered at the subject level. * significant at 10% level, ** significant at 5% level, *** significant at 1% level.

Parameter	Estimate
α	0.413009
	(0.295782)
δ	0.214402
	(0.163174)
β_{length}	0.001061
	(0.002720)
β_{rounds}	0.000256
	(0.000401)
N = 3564	

Table B.2: Results for nonlinear least squares estimation of the Rabin and Vayanos (2010) model for type ‘both’ subjects according to equation 1. Standard errors clustered at the subject level. * significant at 10% level, ** significant at 5% level, *** significant at 1% level.

Parameter	Estimate
α	0.002218
	(0.002965)
δ	1.000000***
	(0.206427)
β_{length}	0.001244
	(0.000813)
β_{rounds}	-0.000106
	(0.000104)
$N = 4752$	

Table B.3: Results for nonlinear least squares estimation of the Rabin and Vayanos (2010) model for type 'rat' subjects according to equation 1. Standard errors clustered at the subject level. * significant at 10% level, ** significant at 5% level, *** significant at 1% level.

B.2 Empirical Entropy

Similarly to section 4.4 in the main text, tables B.4, B.5 and B.6 show regression results for, respectively, all subjects, type 'both' subjects and type 'rat' subjects.

B.3 Generalized Rabin and Vayanos

Similarly to section 4.6, table B.7 reports the results of the generalized Rabin and Vayanos model for all subjects.

	q_0
const	0.0027
	(0.011)
ΔH_0	0.0400
	(0.033)
ΔH_1	0.0731***
	(0.023)
ΔH_2	-0.0341
	(0.026)
length	-0.0003
	(0.002)
rounds	0.00008
	(0.000)
R^2	0.003
N	16,038

Table B.4: Regression results for all subjects. q_0, the dependent variable, is the reported probability of 0 as the next outcome. OLS regression with standard errors clustered at the subject level. * significant at 10% level, ** significant at 5% level, *** significant at 1% level.

	q_0
const	0.0183
	(0.031)
ΔH_0	0.0357
	(0.047)
ΔH_1	0.1004
	(0.062)
ΔH_2	-0.0869
	(0.067)
length	-0.0008
	(0.004)
rounds	0.00000
	(0.001)
R^2	0.002
N	3,564

Table B.5: Regression results for type ‘both’ subjects. q_0, the dependent variable, is the reported probability of 0 as the next outcome. OLS regression with standard errors clustered at the subject level. * significant at 10% level, ** significant at 5% level, *** significant at 1% level.

	q_0
const	0.0058
	(0.006)
ΔH_0	0.0102
	(0.011)
ΔH_1	0.0085
	(0.010)
ΔH_2	-0.0078
	(0.014)
length	0.0006
	(0.001)
rounds	-0.0002
	(0.000)
R^2	0.001
N	4,752

Table B.6: Regression results for type 'rat' subjects. q_0, the dependent variable, is the reported probability of 0 as the next outcome. OLS regression with standard errors clustered at the subject level. * significant at 10% level, ** significant at 5% level, *** significant at 1% level.

	q_0
const	0.0054
	(0.012)
pos_1	0.0650***
	(0.009)
pos_2	0.0270***
	(0.007)
pos_3	0.0052
	(0.007)
pos_4	-0.0117*
	(0.007)
pos_5	-0.0062
	(0.007)
pos_6	-0.0129**
	(0.006)
pos_7	-0.0170**
	(0.008)
pos_8	-0.0100
	(0.007)
rounds	0.0001
	(0.000)
size	-0.0006
	(0.002)
R^2	0.026
N	16038

Table B.7: Regression results for all subjects of the generalized Rabin and Vayanos model. q_0, the dependent variable, is the reported probability of 0 as the next outcome. OLS regression with standard errors clustered at the subject level. * significant at 10% level, ** significant at 5% level, *** significant at 1% level.