International Journal of Approximate Reasoning 162 (2023) 109030

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

journal homepage: www.elsevier.com/locate/ijar

Making decisions with evidential probability and objective Bayesian calibration inductive logics

Mantas Radzvilas [a], William Peden [b], Francesco De Pretis [c,*]

[a] Department of Philosophy, University of Konstanz, Germany
[b] Department of Philosophy, Lingnan University, Hong Kong Special Administrative Region
[c] Department of Communication and Economics, University of Modena and Reggio Emilia, Italy

ARTICLE INFO

Article history:
Received 2 May 2023
Received in revised form 28 July 2023
Accepted 12 September 2023
Available online 19 September 2023

Keywords:
Agent-based modelling
Decision under uncertainty
Frequentist statistics
Imprecise probability
Machine learning
Objective Bayesianism

ABSTRACT

Calibration inductive logics are based on accepting estimates of relative frequencies, which are used to generate imprecise probabilities. In turn, these imprecise probabilities are intended to guide beliefs and decisions — a process called "calibration". Two prominent examples are Henry E. Kyburg's system of Evidential Probability and Jon Williamson's version of Objective Bayesianism. There are many unexplored questions about these logics. How well do they perform in the short-run? Under what circumstances do they do better or worse? What is their performance relative to traditional Bayesianism?

In this article, we develop an agent-based model of a classic binomial decision problem, including players based on variations of Evidential Probability and Objective Bayesianism. We compare the performances of these players, including against a benchmark player who uses standard Bayesian inductive logic. We find that the calibrated players can match the performance of the Bayesian player, but only with particular acceptance thresholds and decision rules. Among other points, our discussion raises some challenges for characterising "cautious" reasoning using imprecise probabilities. Thus, we demonstrate a new way of systematically comparing imprecise probability systems, and we conclude that calibration inductive logics are surprisingly promising for making decisions.

1. Introduction

An inductive logic is a formal system that extends deductive logic to include the comparative assessment of deductively invalid arguments. When coupled with a decision theory, an inductive logic provides a formal guide to decisions and beliefs under uncertainty, which can be applied by human reasoners, intelligences, and so on.

There are many inductive logic systems. However, one interesting class, which we shall call "calibration inductive logics", involves accepting hypotheses about relative frequencies. They also provide rules for determining imprecise probabilities for conclusions given these statistical hypotheses. These rules for imprecise probabilities are called "calibration rules" [65]. Finally, these imprecise probabilities can be paired with epistemological or decision-theoretic rules to determine beliefs and decisions.

* Corresponding author.
E-mail address: francesco.depretis@unimore.it (F. De Pretis).

https://doi.org/10.1016/j.ijar.2023.109030

We can estimate relative frequencies using our data and our background knowledge, but our relevant information is typically imprecise — we usually only know that the relative frequency is somewhere in a certain interval. For this reason, calibration inductive logics use imprecise probabilities to represent our information about relative frequencies. These imprecise probabilities from calibration are not the same as "Imprecise Bayesianism", where imprecise probabilities are interpreted as the properties of a set (usually convex) of probability functions that represents an individual or group's beliefs; updating of this set proceeds by a generalisation of Bayesian conditionalization [61,5]. In calibration inductive logics, the imprecise probabilities represent information from an individual's total evidence, rather than a psychological state.

Despite the popularity of Imprecise Bayesianism as an inductive logic based on imprecise probabilities, it faces a number of challenges [7,56,34,15]. Without prejudging these debates, it is worth exploring the alternatives based on calibration, which may avoid the challenges of Imprecise Bayesianism and yet still retain the advantages of an imprecise probability-based inductive logic [31,33,68].

In this article, we focus on two calibration inductive logics: Henry E. Kyburg's system of Evidential Probability and Jon Williamson's version of Objective Bayesianism. Both of these approaches have been the subjects of extensive discussion in the past 15 years [64,65,44,37,45,18,50,40], including a special issue on Kyburg's system [58]. We describe each of them in Section 1, before explaining our study design in Sections 2 and 3. We find clear and consistent results in our comparisons of players and their settings. We describe these results in Section 4. In discussing these results in Section 5, we also use our results to raise some conceptual questions for further research.

1.1. Evidential probability

The first version of Evidential Probability[1] was developed by Kyburg in 1961 [27] but we shall only discuss the last version of his system [38]. The idea is to interpret probability as a formal relation between (a) statements about relative frequencies and (b) statements about events, samples, populations, and so on. Like standard Bayesianism, Kyburg's interpretation involves epistemic concepts, because his aim is to model evidential support in terms of probabilities. The essential idea is that imprecise probabilities should be derived, by a formal procedure, from the information about relative frequencies in (a). However, since we almost always lack enough information about relative frequencies to determine a complete prior distribution over what we are thinking about, evidential probabilities are usually interval-valued. For example, from an agent's past biological studies, they may only know that the proportion of plants with red leaves in autumn in Europe is no more than 50% and at least 5%, but the agent cannot determine a more precise prior from their background botanical information.

Evidential Probability starts from the estimation of relative frequencies. There are many conceptual issues involved in applying such methods to accept the estimates, including Kyburg's famous "Lottery Paradox" [27]. We shall focus just on those issues that are relevant for explaining our tests in this article.

Evidential Probability allows for the use of either frequentist statistical methods or Bayesian statistical methods, depending on the context. Kyburg developed a complex set of rules for determining when frequentist-style or Bayesian-style reasoning is appropriate. Briefly, Evidential Probability typically requires the use of estimates from Bayesian methods, but only if there are not more precise estimates from frequentist methods. When Bayesian estimates are vaguer than the frequentist estimates (the former includes the latter as proper subintervals) then Evidential Probability uses the frequentist estimates. Therefore, Kyburg's system typically requires the use of Bayesian reasoning (conditionalisation) when an agent's information about relative frequencies is rich, but tends to require frequentist statistical reasoning when the agent's information is poor [30].[2] In our tests, agents will lack the frequency-based priors that Evidential Probability requires for Bayesian statistical methods. Instead, in contexts like our tests, Evidential Probability requires the use of confidence interval estimation [35].

To summarise, evidential probabilities are interval-valued, formally determined, and (in the context of this article) correspond to confidence interval estimations of the parameter of interest. Following Kyburg, we shall put the evidential probabilities in the unit interval. Note that an evidential probability like [0.49,0.51] is not an estimate of a probability (as in frequentist statistics) but an interval-valued imprecise probability.

Due to its use of confidence interval estimation, Kyburg's inductive logic requires a step of determining a probability for rejecting and accepting estimates. In the context of our agent-based model, such a "level of acceptance" is just the confidence level. For the sake of familiarity, we shall talk in terms of significance levels. Thus, an acceptance level is defined as $1-\alpha$, where $\alpha \in [0,1]$ is the significance level that a test of a relative frequency (a statistical statement) must exceed in order to be accepted by an agent. The estimate that is accepted in Evidential Probability is always the shortest that can be estimated at a particular significance level [36, p. 261].

How should one choose the significance level? Kyburg, Williamson, and other users of Kyburg's system regard this question as a decision problem, which needs to be answered partly on the basis of our preferences (utilities) and partly

[1] We use capitalised "Evidential Probability" for the inductive logic and the lower-case "evidential probability" for particular probability values in this system.

[2] The complexity of Kyburg's system is due to intermediate cases, plus the challenge of selecting among information about different "reference classes" for determining the evidential probability of a hypothesis relative to some information. We set aside these details, because the reference classes that Kyburg's system recommends in our study match what is intuitive [28,31].

on the basis of (mathematical and epistemological) objective constraints [29,17,62]. Different significance levels affect the speed of acceptance: ceteris paribus, a lower significance level increases the sample size that we need to accept a statistical statement.

The next element for the application of Evidential Probability to a decision problem is a rule for making decisions given utilities and evidential probabilities. Kyburg was open to a range of possible decision rules, which might be chosen on pragmatic grounds, such as computational speed [32, p. 148]. While Kyburg pioneered the application of agent-based modelling in formal epistemology [35], he does not seem to have considered the possibility of using such modelling to choose among decision rules. Our article fills this gap in the literature.

1.2. Objective Bayesianism

In the inductive logic literature, an important system that uses Kyburg's rules for inductive inference is Objective Bayesianism, as developed by Williamson [69]. Williamson employs Evidential Probability as a tool to determine the information about relative frequencies that Objective Bayesians should use to guide their reasoning.[3] We begin by explaining Williamson's interpretation of probability. Evidential probabilities are epistemic: they are quantitative representations of how our evidence constrains what we reasonably believe [36, pp. 201–202]. In contrast, Objective Bayesian probabilities have a mix of evidential and pragmatic justifications [63]. On pragmatic grounds, Williamson adds a rule to transition from evidential probabilities to precise probabilities about "rational" degrees of belief: the Objective Bayesian degree of belief in a hypothesis is given by an entropy-maximising probability function (see Section 2.10) given the constraints from Evidential Probability intervals.[4]

From a decision-theoretic perspective, the key consequence of Williamson's modifications is that an Objective Bayesian always has precise credences, and therefore an Objective Bayesian can use expected payoff maximisation as their decision rule [68]. Williamson argues that Objective Bayesian credences are the best way to achieve this result [63]. However, calibrating by accepting statistical statements is the core of Objective Bayesian updating, rather than standard Bayesian conditionalization using one's priors; Williamson defends this deviation from standard Bayesian epistemology [66]. Like Evidential Probability, Objective Bayesianism is a hybrid of frequentist statistics and Bayesian statistics, but one that is closer to traditional Bayesianism, since the ultimate result is a precise probability representing (arguably ideally rational) psychological states. In our agent-based model, Objective Bayesianism will be represented by a calibration agent called *MaxEnt*, whom we explain later.

2. The agent-based model

We begin by explaining our agent-based model's general conditions, before detailing the decision problem in the model, and we finish by describing the agents in the model. Each agent that we compare is a combination of a decision rule and a significance level α. We shall call these agents "players" because their decision problem will be a (solo) game. Supporters of imprecise probabilities have argued for one or other rule on general theoretical grounds [43,29,55,39,6]. Hence, we decided to investigate the performance of a range of different decision rules. Each agent represents a particular rule, and we investigated each rule with respect to three different values of α. We also included a standard Bayesian agent as a benchmark.

2.1. Decision problem

Our agent-based model is designed for testing players' ability to acquire and use information to make successful decisions. Each test is an iterative sequence of a decision problem — a game — in which players separately observe a sequence of binomial trials, which we shall call "coin tosses", before choosing whether to bet on the outcome of the last toss in the game. The players know that these coin tosses are exchangeable and binomial, but they do not have any prior estimates of the coin's bias. In each game, players go through the following sequence: (a) 4 tosses are observed and added to any observations from previous games; (b) the player makes a decision whether and how to bet on an additional 5th toss; (c) the player observes the 5th toss and adds its outcome to their overall set of observations in the test.

Formally, in each observation, players observe either (1) ω_h, a state where the coin lands heads or (2) ω_t, a state where the coin lands tails. The set of states $\Omega := \{\omega_h, \omega_t\}$ with a typical element ω_i thus contains every possible outcome of an observation. A "history" is a sequence of observations up until a particular point in the test. The "no observation" history is $\tilde{s} := \varnothing$. The set $\mathcal{S} := \Omega^m$ includes the possible histories with observations that can be generated by a finite number $m \geq 1$ of coin tosses. Each $s \in \mathcal{S}$ is a sequence $s := (s_1, \ldots, s_m)$ where, for each $j \in [1, m]$, $s_j \in \Omega$. The set of all possible histories is $S := \left\{\mathcal{S} \cup \left\{\tilde{s}\right\}\right\}$ with a typical element s.

[3] He also makes some modifications to the calibration procedure that are not relevant here [62,65, Chapter 7].

[4] Or the closest approximation to such a function. For example, if the evidential probability of a binomial event given some evidence is the half-closed interval (0.5, 1], then there is no probability that is closest to the entropy-maximising value of 0.5, but there can be more entropic probabilities [67, p. 469].

	ω_h	ω_t
h	$(1-\delta)$	$-\delta$
t	$(\delta-1)$	δ
a	0	0

Fig. 1. Player Payoff Matrix.

The players count the frequencies of heads and tails in their observed histories. Thus, we define a counting function for ω_h as $\kappa : \mathcal{S} \to \mathbb{Z}_{\geq 0}$ such that, for every history $s \in \mathcal{S}$, $\kappa(s) = n\left(\left\{s_j \in s : s_j = \omega_h\right\}\right)$, where $n(\cdot)$ is the cardinality of the set.

Given the observed histories, players have an option to bet on the 5th toss in every game. An action's outcome is defined by a function $\pi : C \times \Omega \to \mathbb{R}$, where $C := \{h, t, a\}$ is the set of possible actions that players can choose, with a typical element c. The function π assigns, to every possible action-state combination $(c, \omega_i) \in C \times \Omega$, a real number $\pi(c, \omega_i) \in \mathbb{R}$ representing the payoff from action $c \in C$ given the state $\omega_i \in \Omega$. Players know, with certainty, the information in **Fig. 1**, which gives the action payoffs associated with every possible outcome of the 5th toss.

$\delta \in [0, 1]$ is a randomly generated value that determines the ticket prices for h and t under winning and losing. We explain the generation of δ in Section 3. A player wins $(1-\delta)$ for choosing h when the 5th toss results in state ω_h and δ for choosing t when the 5th toss results in state ω_t. They suffer a loss of $-\delta$ when they choose h and the 5th toss results in ω_t. They lose $(\delta - 1)$ when they choose t and the 5th toss results in ω_h. Note that abstaining, a, provides a guaranteed payoff of 0.

2.2. The standard Bayesian

As a benchmark player, we included a standard Bayesian player, whom we call *Stan*. We are not attempting an overall epistemological comparison of *Stan* and the calibration players that we describe below, because that would involve many issues beyond the scope of this article. Instead, *Stan* is a point of comparison, because decision theory with Bayesian expected payoffs is relatively well-studied and often used in statistics, economics, and elsewhere.

We define an epistemic model $M_B := \left\{\Omega, \Theta, S, \kappa, p\right\}$, where Ω is the set of states, $\Theta := \left\{x \in \mathbb{R} : x \in [0, 1]\right\}$ is the set of coin biases[5] towards ω_h with a typical element θ, S is the set of possible observation histories, κ is the counting function for ω_h, and $p : S \to \Delta(\Theta)$ is a credence function that assigns, to every history $s \in S$, a probability distribution $p(s) \in \Delta(\Theta)$ on Θ, where $p(\theta \mid s) \in (0, 1)$ is the marginal probability of a coin bias $\theta \in \Theta$ given a history s.

For *Stan's* credences, we used a beta distribution prior. These priors are a very common Bayesian tool for the decision problem that we use in our tests. Beta distribution priors have the useful feature that, with Bayesian updating in our decision problem, their posteriors will also be beta distributions. A beta distribution can be summarised as $\mathrm{B}(\mathrm{a}, \mathrm{b})$, where $\mathrm{a} > 0$ and $\mathrm{b} > 0$ are its shape parameters. *Stan*'s bias towards ω_h is given by a and their bias towards ω_t is given by b. Thus, as a and b tend towards 0, *Stan*'s credences converge more quickly to their sample's frequencies for ω_h and ω_t, while as a and b tend towards infinity, *Stan*'s credences become less responsive to evidence.

A flat prior of $\mathrm{B}(1, 1)$ is a common choice among Bayesian statisticians for problems like our game, where all *Stan* knows is that the game consists of Bernoulli trials (binomial events in an exchangeable sequence) and betting with randomised payoffs. A flat prior updates relatively quickly. A flat prior is also equivocal, which many Bayesian statisticians would regard as required (or at least permissible) in such a game. Therefore, *Stan* will use a flat prior beta distribution.

Stan's evidence consists of observing a sequence of coin tosses. *Stan* updates their credence in each $\theta \in \Theta$ by revising p using Bayes' rule:

$$p(\theta \mid s) = \frac{p\left(\theta \mid \tilde{s}\right) p(s \mid \theta)}{p(s)}, \text{where } p\left(\theta \mid \tilde{s}\right) > 0 \text{ denotes a prior probability of } \theta. \tag{1}$$

Since *Stan* knows that each toss is a Bernoulli trial and the game generates a binomial distribution, Bayes' rule can be reformulated using a counting function κ for each history $s \in \mathcal{S}$ and every coin bias $\theta \in \Theta$ as

$$p(\theta \mid \kappa(s), m) = \frac{p\left(\theta \mid \tilde{s}\right) p(\kappa(s), m \mid \theta)}{p(\kappa(s), m)}, \tag{2}$$

where $p(\kappa(s), m \mid \theta) = \binom{m}{\kappa(s)} \theta^{\kappa(s)} (1-\theta)^{m-\kappa(s)}$. Thus, the posterior probability distribution $p(\kappa(s), m) \in \Delta(\Theta)$ and the prior probability distribution $p\left(\tilde{s}\right) \in \Delta(\Theta)$ are both beta distributions.

[5] For brevity, we shall say "coin bias" rather than "coin toss bias".

The prior credence in any bias can be described as

$$p\left(\theta \mid \mathrm{a}, \mathrm{b}\right) = \frac{\theta^{\mathrm{a}-1}\left(1-\theta\right)^{\mathrm{b}-1}}{\mathrm{B}\left(\mathrm{a}, \mathrm{b}\right)}. \tag{3}$$

Consequently, Bayes' rule for each bias and history is

$$\begin{aligned} p\left(\theta \mid \kappa\left(s\right), m\right) &= \frac{\binom{m}{\kappa\left(s\right)} \theta^{\kappa(s)+\mathrm{a}-1}\left(1-\theta\right)^{m-\kappa(s)+\mathrm{b}-1} \Big/ \mathrm{B}\left(\mathrm{a}, \mathrm{b}\right)}{\int_0^1 \left(\binom{m}{\kappa\left(s\right)} \theta'^{\kappa(s)+\mathrm{a}-1}\left(1-\theta'\right)^{m-\kappa(s)+\mathrm{b}-1} \Big/ \mathrm{B}\left(\mathrm{a}, \mathrm{b}\right)\right) d\theta'} \\ &= \frac{\theta^{\kappa(s)+\mathrm{a}-1}\left(1-\theta\right)^{m-\kappa(s)+\mathrm{b}-1}}{\mathrm{B}\left(\mathrm{a}+\kappa\left(s\right), \mathrm{b}+m-\kappa\left(s\right)\right)}. \end{aligned} \tag{4}$$

Thus, *Stan*'s posterior is another beta distribution characterised by $\mathrm{a}+\kappa\left(s\right)$ and $\mathrm{b}+m-\kappa\left(s\right)$.

For *Stan*'s prior in ω_h, we define an aggregate belief function $\psi^p : S \to \Delta\left(\Omega\right)$, such that for any $\tilde{s}$, the conditional belief in ω_h is

$$\psi^p\left(\omega_h \mid \tilde{s}\right) = \int_0^1 \theta\, p\left(\theta\right) d\theta = \frac{\mathrm{a}}{\mathrm{a}+\mathrm{b}}, \tag{5}$$

while we define the conditional belief in ω_t as $\psi^p\left(\omega_t \mid \tilde{s}\right) = 1 - \psi^p\left(\omega_h \mid \tilde{s}\right)$. *Stan*'s posterior belief in ω_h given $s \in \mathcal{S}$ is

$$\psi^p\left(\omega_h \mid s\right) = \int_0^1 \theta\, p\left(\theta \mid \kappa\left(s\right), m\right) d\theta = \frac{\mathrm{a}+\kappa\left(s\right)}{\left(\mathrm{a}+\kappa\left(s\right)\right)+\left(\mathrm{b}+m-\kappa\left(s\right)\right)}, \tag{6}$$

while their posterior belief in ω_t given that history s is $\psi^p\left(\omega_t \mid s\right) = 1 - \psi^p\left(\omega_h \mid s\right)$.

We define *Stan*'s expectation-based reasoning about actions using a model $D_B := \left\{C, S, \pi, \psi^p\right\}$, where C is the set of possible actions, S is the set of possible observation histories, π is the payoff function, and ψ^p is *Stan*'s aggregate belief function. For any history $s \in S$, the expected payoff from some action $c \in C$ is

$$E\left[c \mid \psi^p, s\right] = \pi\left(c, \omega_h\right)\psi^p\left(\omega_h \mid s\right) + \pi\left(c, \omega_t\right)\psi^p\left(\omega_t \mid s\right). \tag{7}$$

It follows that *Stan* always has a specific real number representing the expected payoff associated with each action. The choice of an action associated with the highest expected payoff is considered to be an optimal (i.e. expected payoff maximising) choice, and thus choosing such an action is rational according to standard Bayesian decision theory [52]. However, the optimal choice may not be unique: there may be more than one action that maximises *Stan*'s expected payoff. The non-uniqueness of the optimal choice is a common result in many contexts of application of Bayesian decision theory (for a technical discussion of applications of Bayesian decision theory models to different types of decision problems, see, for example, [14,26,49,4]). In such situations, a player's final choice of an action remains undetermined by the Bayesian decision theory model: according to the expected utility maximisation criterion, the choice of every action that maximises expected payoffs is a permissible, and therefore possible, choice.

Optimal choices can also be non-unique according to other decision rules considered in this study.[6] In such cases, the criterion cannot be used to determine the final choice of such player, since, according to the decision criterion, each choice that satisfies it is a permissible, and therefore possible, choice.

Many applications of decision theory that require fully determinate choices deal with the non-uniqueness problem by introducing tie-breakers: secondary decision criteria that the decision-maker uses to select among actions that are optimal according to the primary decision criterion. The player who uses additional tie-breakers essentially uses a compound decision rule that involves two or more decision criteria. This approach, however, cannot be applied for our purposes. The goal of this study is to compare the performance of individual decision criteria. To do this, we need to obtain a record of the consequences of the choices that each individual decision criterion prescribes in isolation from other decision criteria. An introduction of tie-breakers would generate compound decision rules that would make it impossible to determine the performance of compound decision rules' components. Moreover, the use of some of the tie-breakers suggested in the literature would simply generate compound decision rules that combine two or more individual decision rules that we are aiming to compare in this study [42]. Although such a study could inform the design of successful and practically applicable decision rules, the introduction of compound decision rules would not attain this study's objectives.

[6] In this particular study, this occurs only for one player, *Dominance*, but other players can have non-uniqueness in other decision problems.

Therefore, we deal with the non-uniqueness problem by assuming that each player responds to non-uniqueness by uniformly randomizing among actions that are optimal according to the primary decision criterion. This assumption allows us to represent the player as a decision-maker who uses only one decision criterion to evaluate the actions, and thus is strictly indifferent between actions that satisfy it. The uniform probability distribution puts equal weight on each action that satisfies the decision criterion, and thus allows us to fully implement the assumption that the player regards each action satisfying their rule as equivalent.

In addition, the uniform randomization assumption allows us to model the non-uniqueness situations as having a "neutral" effect on player's overall performance. Suppose that player's decision criterion selects a choice set $\{h,t\}$. The player then chooses a randomized action $\sigma \in \Delta(\{h,t\})$, where $\Delta(\{h,t\})$ is the set of all the possible probability distributions on the set $\{h,t\}$, that assigns probability $\sigma(h) = \frac{1}{2}$ to action h and probability $\sigma(t) = \frac{1}{2}$ to action t. Given any probability distribution $\varrho \in \Delta(\Omega)$ on Ω, where $\varrho(\omega_h) \in [0,1]$, a player's *ex ante* expected payoff from σ is $\frac{1}{2}\left(\varrho(\omega_h)(1-\delta) + \left(1-\varrho(\omega_h)\right)(-\delta)\right) + \frac{1}{2}\left(\varrho(\omega_h)(\delta-1) + \left(1-\varrho(\omega_h)\right)(\delta)\right) = \frac{1}{2}\left(\varrho(\omega_h) - \delta\right) + \frac{1}{2}\left(-\varrho(\omega_h) + \delta\right) = 0$. Thus, a uniform randomization over $\{h,t\}$ has an *ex ante* expected payoff of 0. The same results can be shown to hold for a uniform randomization on $\{h,t,a\}$. Given a randomized action $\varsigma \in \Delta(C)$, such that $\varsigma(h) = \varsigma(t) = \varsigma(a) = \frac{1}{3}$, the *ex ante* expected payoff from ς given ϱ is $\frac{1}{3}\left(\varrho(\omega_h) - \delta\right) + \frac{1}{3}\left(-\varrho(\omega_h) + \delta\right) + \frac{1}{3}(0) = 0$. Thus, if the player adopts the randomization principle for each decision situation where that player faces a non-uniqueness problem, then, no matter which of the possible coin biases is used, the overall *ex ante* expected payoff from the consistent application of the randomization procedure is 0.

Returning to the particular case of *Stan*, note how they possess precise expectations and payoffs. Hence, they can use payoff maximisation to make their decisions. *Stan* randomly chooses an action among those with the highest expected payoff given their credences, meaning that they make a selection from the set according to a uniform probability distribution. The set of such actions can be defined as

$$\mathrm{B}_{p,s} := \left\{ c \in C : c \in \arg\max_{c' \in C} \left(E\left[c' \mid \psi^p, s \right] \right) \right\}. \tag{8}$$

2.3. Calibration players

In this subsection, we shall describe (1) the updating procedure used by the calibration players and (2) their decision rules. The calibration players include the combinations of Evidential Probability with different decision rules, including a player who can also be interpreted as an Objective Bayesian. To describe the players, we shall explain a generic player called *Calibration*, who becomes a specific player in our agent-based model when they are paired with a specific decision rule. Initially, *Calibration* lacks any information about the relative frequency of heads tosses in the reference class of coin tosses. They just know that the coin tosses are exchangeable and that a toss can only land heads or tails.

Calibration's reasoning is represented using a model $M_F := \{\Omega, \mathrm{A}, \kappa, S, \lambda\}$, where Ω is the set of states, A is the set of considered significance levels with a typical element α, κ is the counting function for ω_h, S is the set of possible histories, and $\varphi : \mathrm{A} \times S \to \mathcal{P}([0,1])$ is the function that assigns, to every significance level-history pair $(\alpha, s) \in \mathrm{A} \times S$, a Clopper-Pearson interval $\varphi(\alpha, s) := (\phi_L, \phi_U) \in \mathcal{P}([0,1])$. The Clopper-Pearson interval is a commonly used method for determining a binomial confidence interval. It guarantees a coverage level that is never lower than the nominal level $1-\alpha$. Its primary technical advantage is that, due to relationship between the binomial distribution and beta distributions, the Clopper-Pearson interval can be easily derived from beta distributions with the counting function $\kappa(\cdot)$ defined in section 2.1. Moreover, unlike some confidence interval estimation methods for binomial distributions (such as the normal approximation or the Jeffreys approximation) the Clopper-Pearson method can be reliably used even given very small samples, which occur in the early parts of our tests.

The lower bound $\phi_L \in [0,1]$ and the upper bound $\phi_U \geq \phi_L$ of this interval can be represented in terms of beta distribution quantiles:

$$\phi_L = \mathrm{B}\left(\frac{\alpha}{2}; \kappa(s), m - \kappa(s) + 1\right); \tag{9}$$

$$\phi_U = \mathrm{B}\left(1 - \frac{\alpha}{2}; \kappa(s) + 1, m - \kappa(s)\right). \tag{10}$$

For a significance level α and a history of m coin tosses $s \in S$ with $\kappa(s) \geq 0$ "heads", *Calibration* estimates the actual coin bias to be within the Clopper-Pearson interval $\varphi(\alpha, s)$. Hence, they reject any values $\phi \notin \varphi(\alpha, s)$. For an Evidential Probabilist, in a scenario like the coin tossing problem where any other statistical statement would be eliminated, the confidence interval is the (imprecise) probability that the coin toss lands on heads.[7] For an Objective Bayesian, this evidential probability provides constraints on the application of entropy-maximisation to generate the (precise) Objective Bayesian probability.[8]

[7] Note the difference with standard frequentist statistical reasoning, where the confidence interval is not a probability.

[8] In his illustrations of Objective Bayesianism, Williamson has used confidence interval estimation with the normal approximation given relatively large samples [69]. Since the samples in our tests can be too small to use the normal approximation of the binomial, our players use Clopper-Pearson methods to directly estimate the binomial confidence level.

Calibration's expectation-based reasoning about actions can be represented with a model $D_F := \{\Omega, \mathrm{A}, S, C, \pi, \mathcal{E}\}$, where $\mathcal{E} : C \times \mathrm{A} \times S \to \mathcal{P}(\mathbb{R})$ is a function that assigns, to every action-significance level-history combination $(c, \alpha, s) \in C \times \mathrm{A} \times S$, an expected payoff vector $\mathcal{E}(c, \alpha, s) := (E[c \mid \phi_l], \ldots, E[c \mid \phi_u])$, where $\phi_l = \min(\varphi(\alpha, s))$, $\phi_u = \max(\varphi(\alpha, s))$ and each expectation $E[c \mid \phi] := \phi\pi(c, \omega_h) + (1 - \phi)\pi(c, \omega_t)$ for every $\phi \in \varphi(\alpha, s)$.

For many of the decision rules that define specific calibration players, two important terms will be the minimum and maximum expectations of an action given a confidence interval. Given an action-significance level-history combination $(c, \alpha, s) \in C \times \mathrm{A} \times S$, we define the minimum expectation of an action c as

$$E^{min}_{c|\alpha,s} := \min_{E[c|\phi] \in \mathcal{E}(c,\alpha,s)} (E[c \mid \phi]), \tag{11}$$

and the maximum expectation as

$$E^{max}_{c|\alpha,s} := \max_{E[c|\phi] \in \mathcal{E}(c,\alpha,s)} (E[c \mid \phi]). \tag{12}$$

Note how $E^{min}_{a|\alpha,s} = E^{max}_{a|\alpha,s} = 0$ for each $(\alpha, s) \in \mathrm{A} \times S$, because a has a guaranteed payoff of zero.

2.4. Dominance

We shall now define the individual players based on the *Calibration* template. The first player, *Dominance*, uses the Interval Dominance rule, sometimes known as the "non-dominated set rule" [9, p. 15]. If an action c's maximum expected payoff is strictly less than the minimum expected payoff of one or more alternatives, then c is "dominated". Interval Dominance permits any non-dominated action. *Dominance* randomises among the set of non-dominated actions.

We define the set of non-dominated actions, given any $(\alpha, s) \in \mathrm{A} \times S$ as

$$\mathrm{D}_{\alpha,s} := \left\{ c \in C : E^{max}_{c|\alpha,s} \geq E^{min}_{c'|\alpha,s}], \text{ for every } c' \in C \right\}. \tag{13}$$

Dominance randomly (meaning according to a uniform distribution) chooses an action c from $\mathrm{D}_{\alpha,s}$.

2.5. E-Admissibility

The next rule was developed as a decision rule for imprecise probabilities by Isaac Levi [41]. We shall call this player *E-Admissibility*, after Levi's name for his rule. It has not previously been considered as a rule for calibrated inductive logics, but it can be adapted to any type of imprecise probabilities, including Kyburg's Evidential Probability system.

A set of actions where each action in the set maximizes *E-Admissibility's* expected payoff with some element of $\varphi(\alpha, s)$ can be defined as

$$\mathrm{E}_{\alpha,s} := \left\{ c \in C : \exists \phi \in \varphi(a, s), c \in \arg\max_{c' \in C} \left(E\left[c' \mid \phi\right] \right) \right\}. \tag{14}$$

E-Admissibility randomly chooses an action c from $\mathrm{E}_{\alpha,s}$.

2.6. Maximin

The next rule is an adaptation of maximin decision theory to imprecise probabilities. In the context of Imprecise Bayesianism, it has been called Γ-Maximin [55], but for simplicity's sake we shall just call it Maximin. This rule instructs us to maximise the minimum expected payoff of our actions.

A set of actions that maximize the minimum expectation given a significance level α and a history s can be defined as

$$\mathrm{M}_{\alpha,s} := \left\{ c \in C : E^{min}_{c|\alpha,s} \geq E^{min}_{c'|\alpha,s}, \text{ for every } c' \in C \right\}. \tag{15}$$

Thus, each c in the set is minimum payoff-maximising according to some coin bias that is consistent with *Maximin*'s estimated interval in a particular game. The player *Maximin* randomly chooses an action c from $\mathrm{M}_{\alpha,s}$.

2.7. Regret

The Minimax Regret rule was developed to address a criticism of Maximin rules: the latter is insensitive to opportunity costs [52,57,6].[9] The *Regret* player makes decisions based on a "regret number" for each action, which is the highest

[9] The opportunity cost of a choice is the value of the best possible alternatives. In contrast, accounting cost is the absolute net profit. For example, if the loss from making a particular choice is $20, but your best alternative was losing $15, then the loss in terms of accounting costs was $20. However, your opportunity cost was $15, then you lost just $5 in opportunity cost terms. If your best alternative was losing $20, then you lost nothing in terms of opportunity costs.

opportunity cost of the action (in terms of expected payoffs) for any coin bias that is compatible with *Regret*'s confidence interval. Put another way, an action's regret number represents the greatest difference, according to some coin bias that is consistent with the confidence interval, between (1) the highest expected payoff associated with any possible action, and (2) the actual expected payoff of that action. In a particular game, *Regret* chooses an action with a minimal regret number, given their confidence interval estimate. They choose an action that minimises the maximum regret number.

The set of regret-minimising choices can be defined with a regret function $r : C \times \mathrm{A} \times S \to \mathbb{R}$ that assigns, to each combination $(c, \alpha, s) \in C \times \mathrm{A} \times S$, a regret value

$$r(c, \alpha, s) := \max_{\phi \in \varphi(\alpha, s)} \left(\max_{c' \in C} \left(E\left[c' \mid \phi\right]\right) - E\left[c \mid \phi\right] \right). \tag{16}$$

We define the set of regret-minimising choices as

$$\mathrm{R}_{\alpha,s} := \left\{ c \in C : c \in \arg\min_{c' \in C} \left(r\left(c', \alpha, s\right)\right) \right\}. \tag{17}$$

Regret randomly chooses an action c from $\mathrm{R}_{\alpha,s}$.

2.8. The Hurwicz players

The next rule, developed by Leonid Hurwicz, uses a constant parameter $\gamma \in [0, 1]$ to quantify different degrees of cautiousness [20]. When $\gamma = 1$, a Hurwicz player only makes decisions using their minimum expectations. Similarly, when $\gamma = 0$, then a Hurwicz player only takes their maximum expectations into account when making decisions.

In our model, we examine γ values of 0.25, 0.5, and 0.75, because it is with such middling values that the Hurwicz approach is distinct from other rules. The *Pessimist* player has $\gamma = 0.75$. The *Intermediate* player has $\gamma = 0.5$. The *Optimist* player has $\gamma = 0.25$.

A set of actions that maximize the Hurwicz measure given some combination (γ, α, s) can be defined as

$$\mathrm{H}_{\gamma,\alpha,s} := \left\{ c \in C : c \in \arg\max_{c' \in C} \left(\gamma E^{min}_{c'|\alpha,s} + (1 - \gamma) E^{max}_{c'|\alpha,s} \right) \right\}. \tag{18}$$

The Hurwicz players randomly choose an action c from $\mathrm{H}_{\gamma,\alpha,s}$.

2.9. Opportunity risk optimisation

This decision rule was first suggested by Daniel Ellsberg [11, p. 664], building on earlier research [19]. Ellsberg does not give it a name. However, it balances two goals: (1) maximising payoff opportunities and (2) avoiding extreme risks, so we call it *ORO*: Opportunity-Risk Optimisation. *ORO* calculates a second-order parameter using two values: (1) the average expected payoff given their confidence interval, plus (2) the minimum expected payoff for each action. These factors are weighted by *ORO*'s confidence in their confidence interval estimate. We formalise this "confidence" in terms of the length of the interval, as defined below.

For any combination $(c, \alpha, s) \in C \times \mathrm{A} \times S$, the average expected payoff can be defined as

$$E^{avg}_{c|\alpha,s} := \phi^{avg} \pi\left(c, \omega_h\right) + \left(1 - \phi^{avg}\right) \pi\left(c, \omega_t\right), \text{ where } \phi^{avg} := \frac{\phi_l + \phi_u}{2}. \tag{19}$$

The set of actions that are optimal for the ORO player given some $(\alpha, s) \in \mathrm{A} \times S$ can be defined as

$$\mathrm{O}_{\alpha,s} := \left\{ c \in C : c \in \arg\max_{c' \in C} \left((\phi_u - \phi_l) E^{min}_{c'|\alpha,s} + (1 - (\phi_u - \phi_l)) E^{avg}_{c'|\alpha,s} \right) \right\}. \tag{20}$$

The *ORO* player randomly chooses an action c from $\mathrm{O}_{\alpha,s}$.

2.10. MaxEnt

The next rule corresponds to what is required by Williamson's version of Objective Bayesianism, when applied to our problem. It could also be applied by a user of Evidential Probability or a frequentist reasoning under uncertainty, since the entropy-maximising probabilities can be interpreted as auxiliary quantities for making decisions, rather than as rational degrees of belief.[10]

The Maximum Entropy Principle extends the Principle of Indifference – that, in situations of complete ignorance (where we have no information about the relevant relative frequencies) we should generate probabilities by assigning equal probabilities to what we believe to be the fundamental states of the world. The Principle of Indifference is a rule for complete

[10] See also John Maynard Keynes's suggestion [25, p. 214].

ignorance, whereas the Maximum Entropy Principle can be applied outside situations of complete ignorance [21]. In the context of our agent-based model, this means that the player uses an entropy-maximising probability distribution given the constraints provided by their confidence interval estimates.

To define the *MaxEnt* player, we define a probability function $u : \Theta \to \Delta(\Theta)$ which assigns a uniform probability distribution on the set Θ and a belief function $\varpi : \mathcal{P}([0,1]) \to \mathcal{P}([0,1])$ that assigns, to any interval $\varphi(\alpha, s) \in \mathcal{P}([0,1])$, a set of entropy-maximizing beliefs $\varpi(\varphi(\alpha, s)) \subseteq \varphi(\alpha, s)$, where each belief $w \in \varpi(\varphi(\alpha, s))$ is such that

$$w \in \arg \min_{\phi \in \varphi(\alpha,s)} \left| \left(\phi - \int_0^1 \theta u(\theta)\, d\theta \right) \right|. \tag{21}$$

Since $\int_0^1 \theta u(\theta)\, d\theta = \frac{1}{2}$, equation (21) can be simplified to

$$w \in \arg \min_{\phi \in \varphi(\alpha,s)} \left| \phi - \frac{1}{2} \right|. \tag{22}$$

A set of entropy-maximizing actions given some $(\alpha, s) \in \mathrm{A} \times S$ can be defined as

$$\mathrm{W}_{\alpha,s} := \left\{ c \in C : \exists w \in \varpi(\varphi(\alpha, s)), c \in \arg\max_{c' \in C} \left(E\left[c' \mid w \right] \right) \right\}. \tag{23}$$

MaxEnt randomly chooses an action c from $\mathrm{W}_{\alpha,s}$.

2.11. Significance levels

We decided to test a range of significance levels, since a wide range of significance levels are consistent with the calibration inductive logics developed by Kyburg and Williamson. For values of α, we used 0.01, 0.05, and 0.5. The first two correspond to common choices of values in frequentist statistics. The 0.5 value has performed well in earlier studies [35,50].

3. Test methods

Our aim was to use our agent-based methods to test the *Calibration* players (those using frequentist statistical methods) against each other, as well as against our benchmark player *Stan*. In this section, we detail our methods for conducting these tests.

Each "test" has 1000 games. The tosses in each particular test were randomly generated for a particular coin bias. We defined each bias in terms of heads, with 1 being a coin that always lands heads, 0 being a coin that always lands tails, 0.5 being a toss that lands heads 50% in the long-run, and so on. To test players with a broad spread of biases, we tested biases of 0.1, 0.3, 0.5, 0.7, and 0.9.

Using a powerful server, we created a large number of tests (1000 per player setting) to greatly reduce the risk of random errors. The players made all their decisions separately. They did not retain information from previous tests. For the sake of fairness, we used the same randomly generated coin toss outcomes for all players. Since there were 1000 games and 5 tosses per game, there were 5000 tosses per test. Each player also was faced with the same 1000 randomly generated ticket prices in a particular test. We randomly generated the parameter δ using a uniform distribution; these parameter values determined the values in **Fig. 1**. Each test was fair, in the sense that players were evaluated using the same coin toss outcomes and ticket prices. This design made our tests into controlled experiments, where we could vary one factor (a decision rule, significance level, coin bias, and so on) at a time.

Our tests can be regarded as a set of random observations, because each was stochastically independent of the others, while coin tosses and ticket prices were randomly generated, and the player parameters (such as α) were constant for a given test setting. Consequently, we used confidence interval estimation. It is crucial to note that, in some cases, our metrics for comparing players have overlapping confidence intervals. Thus, all of our analyses of relative player performances are always in terms of the *confidence intervals* around averages, not the averages themselves. All our results, provided at the end of the article, were estimated as confidence intervals at the 0.95 confidence level.

We coded ten different scripts using Python 3 (version 3.8.1), with the `statsmodel` econometric and statistical library [53]. The first two scripts generated coin toss outcomes and ticket prices. The other eight scripts generated average monetary net profits for each player (see **Figs. 2 to 9** plus the statistics for the conditions met by players in the tests — see **Table 2** to **Table 9**. All tests were performed on an Ubuntu Linux server powered by a 64 cores (128 threads) Intel Xeon (Phi type) processor with a clock speed of 1.3 GHz, coupled with 128 GB RAM.

3.1. Comparison criteria

We used three comparison tools. First, we compared players' average aggregate performance in terms of the goodness of fit between their decisions and those that they (or any other player) would make if they knew the true bias in a test setting. Second, we used graphical comparisons of the evolution of their average decision-making performance, in terms of average payoffs, over all simulations. Third, we compared calibration players against *Stan* using the aforementioned two criteria.

3.1.1. Aggregate comparisons

To quantitatively compare players' overall performances, we aimed to contrast them with respect to the goodness of fit between their (averaged) choices given a coin bias and the distribution of choices that they (or any other player) would have made if they knew the true coin bias at the start of the test. A measure using such an approach has the advantage of comparing player's performances given particular distributions of coin tosses, rather than simply measuring average payoffs, because the performance we can expect from players depends on the bias and the particular distributions that occurred in our tests. The latter factor is not very important due to the large number of tests we performed, which almost entirely wash out random variations. The bias issue is more serious. For example, it is easier to perform well given a bias of 0.9 or 0.1 than a bias of 0.5. Hence, the true long-run bias gives a natural benchmark for assessing players. Thus, we needed a measure that was sensitive to the degree of difficulty created by different tests. We also wanted the measure to be insensitive to particular payoffs in the tests, because the latter fact was randomly generated and unreflective of player's performances as such.

Ideally, we would simply construct, using the average statistics in our tests, (a) a probability distribution corresponding to a player's choices and (b) a probability distribution for the choices that they would make if they knew the true bias. The latter is the same for each player, in that their decision rules require different choices only because they are using imprecise information from confidence interval estimates of the bias. However, the possibility of randomising or abstaining in our games meant that players' choices could not be straightforwardly represented by a probability distribution. Instead, we created proxy statistics for players' behaviour and adapted the Wasserstein metric to measure the goodness of fit between proxy distributions and the true bias [59,24]. In the rest of this subsection, we explain the details of this measure.

Randomisation is a particular challenge that arose from how some of our players sometimes make choices. How should these randomised choices be weighted? Randomising frequently is clearly better than persistently deviating from the true bias. For example, if the true coin bias is 0.9, then it is better to randomise than to act as if the coin bias is 0.1. On the other hand, it is even better to act insofar as one acts on the assumption that the true coin bias is 0.9, when this reflects the result of plausible update and decision rules. Additionally, a player who is randomising between h and t when the bias is 0.5 should not receive the same score as a player who has learned that the bias is 0.5, because we are measuring their capacity to quickly learn the true bias and use *that* information to make choices, not simply their ability to make successful decisions. Thus, we discounted random choices by $1/z$, where a player is randomising among $z > 1$ choices. This approach has further advantages:

1. The more actions that a player rules out (for example, they randomise between just two actions, rather than three) the higher their score, reflecting their greater ability to use information from their observations.
2. The discount rates for two players who randomise in the same way would be identical, even if one was lucky and guessed the correct answer at a higher frequency. Our study is about players' ability to systematically make better choices, so it is important to remove the effects of lucky randomisation.
3. Discounting in this way is proportionate to the difference between a player's choice to randomise and the optimal choice. To simplify, imagine that our tests consisted of just one game. If the optimal choice in that game is t and a player uniformly randomised between t and h, then in the long-run they would be right in 50% of such tests, which matches the discounting of their choices. Similarly, discounting by $1/z$ would match their performance in making the optimal choice if they randomised between t, h, and a.

However, we recognise that any discounting of randomisation could be controversial. In our article, the relevant of these controversies is mitigated by the fact that our other comparison method (described in the next subsection) does not discount randomisation, as it looks purely at average payoffs. Furthermore, only *Dominance* randomises — see **Tables 3 to 9**.

Another issue was how to weigh abstaining. On one version of our measure, abstaining is costly, because it indicates a slower ability to successfully use information from small samples, which is the central challenge in our tests. However, in the original base game that we used (see **Fig. 1**) we set the payoff from the action a to 0. Thus, abstaining gave neither a profit nor a loss. Similarly we included a reparameterisation of our measure that disregards players' abstaining, so that only the goodness of fit of their choices with ideal choices is included in the reparameterised measure. We also recorded the size of the abstaining factors, in order to enable easier assessments of its effects.[11] Furthermore, we provide the raw statistics for how often players abstained, randomised, and so on. Hence, if readers wish to weigh such decisions differently to match their own theories of rationality, they can re-weigh our results.

[11] We thank an anonymous referee for suggesting a neutral treatment of abstaining.

Table 1
Possible player choice conditions.

Condition	Description
K_1	Directly choosing h.
K_2	Directly choosing t.
K_3	Directly choosing a.
K_4	Randomising between h and t.
K_5	Randomising between h and a.
K_6	Randomising between t and a.
K_7	Randomising between h, t, and a.

To summarise, we quantitatively assessed players by assessing their speed in obtaining and using sample information as if they knew the true bias. To make this comparison, we adapted a standard method for measuring the distance between probability distributions. We measured players over two periods, 250 games and 1000 games. We normalised the measure to provide a real number between 0 and 1. For each recorded measure value, we also included a separation of the abstaining factor and the goodness of fit factor in that value. Below, we explain the formal details of this measure and its reparameterisations.

There are seven possible conditions that can occur, which we delineate in **Table 1**. We define a set of player types $\mathrm{T} :=\{$*Dominance*, *E-Admissibility*, *Maximin*, *Regret*, *Optimist*, *Intermediate*, *Pessimist*, *ORO*, *MaxEnt*, *Stan*$\}$ with a typical element τ, and an ordered set of choice conditions $K := (K_1, K_2, K_3, K_4, K_5, K_6, K_7)$, where each condition $K_l \in K$ is an action from **Table 1**. These actions can be either mixed (randomising) or pure (non-randomising). Next is a condition allocation rule $\iota : \mathrm{T} \to \mathcal{P}(K)$ that assigns, to each player type $\tau \in \mathrm{T}$, a set of conditions $\iota(\tau) \in \mathcal{P}(K)$ such that $\iota(\tau) = (K_1, K_2, K_3)$ if and only if $\tau \in \{$*MaxEnt, Stan*$\}$ and $\iota(\tau) = K$ otherwise. The condition allocation rule functions as a formal way to refer to players in our measure's definition. *Stan* and *MaxEnt* are treated separately, because their rules mean that they will never randomise or abstain in our decision problem. By contrast, with suitable payoffs and settings, the other players' decision rules could be compatible with randomising.

The cardinality of the sets $\iota(\tau) \in \mathcal{P}(K)$ can be defined as $n(\iota(\tau)) \in \{3, 7\}$. This enables us to define a function $\mu : C \times \mathrm{T} \to [0,1]^{n(\iota(\tau))}$ that assigns, to every action-player type combination $(c, \tau) \in C \times \mathrm{T}$, a vector of weights $\mu(c,\tau) := \left(\mu_1(c,\tau), \ldots, \mu_{n(\iota(\tau))}(c,\tau)\right)$, where $\mu_l(c,\tau) \in [0,1]$ is the weight of a condition $K_l \in \iota(\tau)$ given an action $c \in C$. The function μ is such that the choices and their circumstances are weighted as follows:

$$\mu(h,\tau) = \begin{cases} (1,0,0) \text{ if and only if } \tau \in \{\textit{MaxEnt, Stan}\}; \\ \left(1,0,0,\tfrac{1}{2},0,\tfrac{1}{2},\tfrac{1}{3}\right) \text{ otherwise.} \end{cases} \tag{24}$$

$$\mu(t,\tau) = \begin{cases} (0,1,0) \text{ if and only if } \tau \in \{\textit{MaxEnt, Stan}\}; \\ \left(0,1,0,\tfrac{1}{2},\tfrac{1}{2},0,\tfrac{1}{3}\right) \text{ otherwise.} \end{cases} \tag{25}$$

$$\mu(a,\tau) = \begin{cases} (0,0,1) \text{ if and only if } \tau \in \{\textit{MaxEnt, Stan}\}; \\ \left(0,0,1,0,\tfrac{1}{2},\tfrac{1}{2},\tfrac{1}{3}\right) \text{ otherwise.} \end{cases} \tag{26}$$

We examined two periods: 250 games and 1000 games. To refer to these periods, we define a set $\Xi := \{250, 1000\}$ with a typical element ξ. Our measure also takes inputs from the statistics of the average number of times that a player met a condition. The data from our tests provides the statistics for the members of a set $\chi^{\tau}_{\theta,\xi} := \left\{\chi^{\tau}_{h|\theta,\xi}, \chi^{\tau}_{t|\theta,\xi}, \chi^{\tau}_{a|\theta,\xi}\right\}$, where each element $\chi^{\tau}_{c|\theta,\xi} \in [0,1]^{n(\iota(\tau))}$ is a vector reporting the mean number of times that a given condition K_l was met in the appropriate set $\iota(\tau)$, given a player type τ, action c, and period ξ. These sets provide the inputs for the performance measure. These are weighted using μ. The data that we used for $\chi^{\tau}_{\theta,\xi}$ are detailed in **Tables 2 to 9**.

Our performance measure is a function that uses the aforementioned terms and generates a real number between 0 and 1 to assess player's aggregate performances. We define the function as $\lambda : \Theta \times \mathrm{T} \times \Xi \to [0,1]$ that assigns, to every coin bias-player type-period of games combination $(\theta, \tau, \xi) \in \Theta \times \mathrm{T} \times \Xi$, a real number $\lambda(\theta,\tau,\xi) \in [0,1]$, such that

$$\lambda(\theta,\tau,\xi) := \sqrt{\underbrace{1 - \left(\mu(a,\tau) \bullet \chi^{\tau}_{a|\theta,\xi}\right)\xi^{-1}}_{\text{Abstaining factor}}} \quad .$$

$$\sqrt{\overbrace{1-\left(2^{-1}-\left|2^{-1}-\theta\right|\right)}^{\text{Normalizing and scaling terms}}\left[\underbrace{\left|\theta-\left(\mu\left(h,\tau\right)\bullet\chi^{\tau}_{h|\theta,\xi}\right)\xi^{-1}\right|\theta^{-1}}_{\text{Goodness of fit in predicting the frequency of heads}}+\overbrace{\left|1-\theta-\left(\mu\left(t,\tau\right)\bullet\chi^{\tau}_{t|\theta,\xi}\right)\xi^{-1}\right|(1-\theta)^{-1}}^{\text{Goodness of fit in predicting the frequency of tails}}\right]}. \quad (27)$$

In this formula, the symbol $\bullet$ denotes the scalar product between vectors. To account for the potential action a (abstain) among players, the goodness of fit must be separately defined for both heads and tails. In other words, the goodness of fit in predicting the frequency of heads can't be retrieved directly by only observing the goodness of fit in predicting the frequency of tails, given the possibility of abstention. Furthermore, certain parts of the formula have been designed to ensure that the resulting measure falls within the familiar unit interval, while the use of absolute values helps prevent cancellations.

To disaggregate the two different information sources that contribute to this performance measure, we have re-parameterized the previous formula by introducing Λ_1 and Λ_2, two terms related to the abstain and the goodness of fit factors defined above:

$$\lambda\left(\theta,\tau,\xi\right):=\sqrt{1-\Lambda_1}\cdot\sqrt{\Lambda_2}, \text{where } \left(\Lambda_1,\Lambda_2\right)\in[0,1]^2. \quad (28)$$

The parameter Λ_1 measure the influence of the abstaining factor by compacting the information in **Tables 2 to 9**. It equals 0 for those players who do not abstain; it increases as the relative frequency of abstention rises. The parameter Λ_2 provides a measure for aggregate comparisons that disregards the abstaining factor, and thus measures just players' goodness of fit performance. We report the values of λ and (Λ_1,Λ_2) in **Tables 10 to 13**.

3.1.2. Graphical comparisons

To examine the evolutionary path of players' performances, we constructed graphs of their average performances (in terms of payoffs) over the games. We report these results in **Figs. 2 to 9**. Each graph is for a particular player, coin bias, and significance level. To plot the graph for each such combination of settings, we used our computation of the average performance over 1000 tests of that player for each game. The zero point in each chart is the average of all players' payoff performances given a coin bias of 0.5, which provides a convenient common scale.

Our evolutionary comparison graphs provide additional information to our aggregate comparisons. Firstly, they show how a player's average performance evolved in a particular test setting. Secondly, the graphs directly test their ability to maximise payoffs in the games, whereas our aggregate comparisons test players ability to approximate the true bias in the short run. Thirdly, the graphs enable disaggregated assessments of players' performances, so we can examine any differences in particular subperiods of a test.

3.1.3. Comparisons with Stan

As a benchmark, we used *Stan*, who has performed well in similar earlier studies that applied agent-based modelling to the evaluation of inductive logics; they performed at least as well as any other player in comparable decision problems [35,50]. Hence, if a player can match or approximate *Stan*'s performance, they can perform as well as any other studied player in such tests, which is a strong point in their favour. Conversely, if *Stan* outperforms a player by a large margin, then that is useful information for assessing whether to use that player's update and decision rules in areas like AI, policy evaluation, or financial-decision making. Thus, in both our graphical comparisons and our aggregate comparisons, we include the results for *Stan*, and we compare the calibration players to this *Stan*'s benchmark performance.

4. Results

In this section, we describe the results of both the aggregate comparisons and the graphical comparisons. We begin by making some general points about players' performances, before discussing specific results and their comparisons. We finish by making comparisons with *Stan*.

As expected, players did better as the tests proceeded — their performance were better as their acquired more data. This provides a useful adequacy check for our tests, since confidence interval estimation should tend to result in more accurate estimates (and hence better average performances) as the sample sizes increase. In contrast, the relationship between aggregate performances and coin biases varied among the players, as we shall detail.

For both the aggregate comparisons and the graphic comparisons, players generally did better as α was higher. Thus, $\alpha=0.05$ led to better performances than $\alpha=0.01$, while $\alpha=0.5$ led to better performances than $\alpha=0.05$. This result may seem surprising, since the lower values of α are more typical in scientific practice, but it mirrors results in similar earlier research [35,50] where players using calibration inductive logics performed better with lower values of α in this decision problem. The explanation is that in this game there is a high value, in terms of opportunity cost, for quickly detecting coin biases. Obviously, the estimates from small samples are less reliable than with larger samples, but they still tend to provide some valuable information, and generally there are payoff rewards for using this information. Hence, insofar as calibration players made more use of these small samples, they did better in this decision problem.

Moreover, while $\alpha = 0.5$ is not reliable according to usual scientific standards, it does have a rational interpretation in Kyburg's Evidential Probability and Williamson's Objective Bayesianism. It means that if the error rate in the significance level is the appropriate statistical information for determining the evidential probability, then the confidence interval estimate is "more probable than not", in the sense that the hypothesis that the confidence interval includes the true coin bias is an imprecise (evidential) probability with a lower bound of 0.5. Additionally, neither system commits us to the view that there is a universal appropriate standard of acceptance. According to Kyburg and Williamson, the level of acceptance can vary with the context [29,65]. Hence, it is compatible with their views that $\alpha = 0.5$ is a possible standard of acceptance in decision problems like the game we study, whereas in scientific contexts like statistical testing of hypotheses a value of 0.05, 0.01, or even lower could be appropriate. In general, we stress that our results do not have the absurd implication that scientists in general should use confidence intervals of 0.5 in statistical estimations. Instead, what we have found is that calibration inductive logics perform better with this standard in this particular type of decision problem.

The size of differences between players' performances varied with α. However, the ordering of differences was the same according to our main aggregate comparison measure λ, as well as our graphical comparisons. Thus, the best performing players with $\alpha = 0.01$ were also the best performing players with $\alpha = 0.5$.

4.1. Aggregate comparisons

Three calibration players made the same decisions and also performed the best on our aggregate comparisons, so we categorise these together as the *Leaders*. These players were *E-Admissibility*, *Optimist*, and *Intermediate*. We report our aggregate comparison results in **Tables 10 to 13**. We shall begin with results for λ in periods of 250 games. We shall only discuss periods of 250 games rather than 1000 games, because it was in the shorter periods that differences between players were most apparent.

In our aggregate comparisons, the *Leaders* performed very well. Unexpectedly, the *Leaders*' performance was best when the bias was 0.5. In fact, under that condition, the difference could only be detected at 2 or more decimal places. The reason is that, on average, their confidence interval estimates will be symmetric around the true bias when the latter is 0.5. Thus, they rapidly approximate the rational decisions given knowledge that the bias is 0.5, and therefore perform better according to λ.

MaxEnt and *Dominance* were the next two best performing players on our aggregate comparisons. The extreme similarity of their performances is interesting, because they are based on two fundamentally different decision rules. In our decision problem, *MaxEnt* initially assigns a probability of 0.5 to the hypothesis that the next coin toss will land heads; they only shift away from this assumption insofar as it is inconsistent with their confidence interval estimates. They never randomise in our decision problem. In contrast, *Dominance* chooses the uniquely dominant action if it exists, and randomises whenever there are two or more undominated actions. However, while these two players do about equally well on average over all biases, they have opposite patterns for particular biases, due to the details of their decision rules.

We shall now explain the causes of those differences. Unsurprisingly, *MaxEnt* does best with a bias of 0.5, since they begin with a prior that matches this value. However, they did better with extreme biases of 0.1 or 0.9 than 0.3 and 0.7. The explanation of the latter pattern is that with an extreme bias, *MaxEnt* is more quickly pulled away from their misleading initial prior, because the sample variance is lower — a bias of 0.1 or 0.9 produces representative samples at a higher rate than 0.3 or 0.7. In contrast, *Dominance* does better insofar as the bias was more extreme, because this provided them with a higher frequency of uniform samples, enabling more precise confidence interval estimates. When *Dominance* lacked uniform samples, they had a greater tendency to randomise between actions; see **Tables 4 to 9** for the precise averages. When randomising, *Dominance* abstains in 1/3 of games in the long-run. Abstaining in these games did not cause an accounting loss, since the payoff of a is always 0, but abstaining did create an opportunity cost relative to the *Leaders*. Thus, *Dominance* underperformed due to abstaining, whereas *MaxEnt* underperformed due to their slow (in comparison to *Stan*, who also starts with an equivocal prior) revision of their prior.

Regret performed the same as *Dominance*, in that they both did better with extreme biases. However, the cause of *Regret*'s slight underperformance was different. As **Tables 4 to 9** show, *Regret* directly chose to abstain in some games. In contrast, *Dominance* only chose a when randomising between h, t, and a. Therefore, by different routes, these players failed to make optimal use of their information. This underperformance occurred because of their decision rules rather than their update rule or significance level, because all the calibration players in our study update in the same way.

The most severe underperformances were by *ORO* and *Maximin*. Both of these players had the same problems, but *Maximin* had marginally greater losses. These two players choose a at the highest rates than other players. Since their expected payoffs are interval-valued, it is possible for both h and t to have low maximal minimum expected payoffs. This is more likely in earlier games, when their confidence intervals are wide, causing wide expected payoff intervals. *ORO* chooses a less often than *Maximin* because *ORO* only uses Maximin-type reasoning as part of its decision rule.

Like *ORO*, *Pessimist* has some fundamental similarities to *Maximin*. The latter is equivalent to a Hurwicz player with $\alpha = 1$, whereas *Pessimist* is a Hurwicz player with $\alpha = 0.75$. *ORO*, *Maximin*, *Regret*, and *Pessimist* were the only players to directly choose a. However, since *Pessimist* chooses a at roughly half the rate of *Maximin*, their performance was statistically significantly better.

To summarise these results for the λ measure, players did better insofar as they abstained less often, avoided unrobust priors, and used a high value of α. Abstaining in earlier games would not cause an underperformance if it were not possible

for players to systematically identify profit opportunities in these games. However, players like *Leaders* can systematically identify these profit opportunities, and thus there is a cost for calibration reasoners to use the rules that sometimes select a in this decision problem.

We now turn to Λ_2 (the aggregate comparison measure that disregards the abstaining factor) and again describe results for periods of 250 games. Given α = 0.01, all players performed roughly equally well given a bias of 0.9. They also performed roughly equally well given a bias of 0.1, except *Maximin*. However, the *Leaders* retained their position as the top performing players under other biases, with the exception of *MaxEnt* matching them given a bias of 0.5.

There were similar results for α = 0.05. The difference was that *ORO* fell slightly below the *Leaders*' performances when the bias was 0.1 and *Maximin* likewise when the bias was 0.9. However, these differences between *Maximin* and *ORO* were very marginal. As with α = 0.01, the only players who were consistent top performers under all coin biases were the *Leaders*.

Finally, given α = 0.5, the only change in the ordering of performances relative to those using λ was that *Dominance*, *Regret*, and *Pessimist* matched the *Leaders* when the bias was 0.1 or 0.9. However, this improvement was not maintained under other coin biases.

Therefore, the results for Λ_2 reiterate the main result when using λ as the aggregate comparison measure. Some calibration players can match *Leaders* under certain conditions. However, no calibration player could consistently match their performance.

4.2. Graphical comparisons

We begin with a few general points. For the value of α, the graphical comparisons corroborated the pattern from the aggregate comparisons: higher values of α correlated to a better performance. Additionally, the ordering of players' performances remained stable over different values of α and coin biases.

We now turn to specific groups of players. The best performers according to our graphs were the *Leaders*, *Regret*, and *ORO*. These players performed best regardless of the coin bias. While these players made different decisions, the differences were not important enough to create an observable difference in the graphs. However, recall that our aggregate comparisons showed an advantage for the *Leaders* over *Regret* and *ORO*.

MaxEnt equalled these players when the bias was 0.5, but did as poorly as any player with a bias of 0.1 or 0.9. When the bias was 0.3 or 0.7, *MaxEnt*'s performance was contingent on the level of α. Given $\alpha = 0.5$, *MaxEnt* performed about as well as any player. However, with lower values of α, there was an observable lag in *MaxEnt*'s convergence (see **Fig. 6**). This pattern occurred because *MaxEnt* does better given biases other than 0.5 if they more quickly revise their initial prior. A higher α value increases the speed of *MaxEnt*'s revision and thereby results in better average performances.

The *Pessimist* player approximately matched the top performers when the bias was 0.1, 0.3, 0.7, or 0.9. However, when the bias was 0.5, they had a slightly slower convergence. This tendency can be explained by *Pessimist*'s abstaining; the cessation of this behaviour is slower with a 0.5 bias, due to the higher sample variance. For the same reason, their underperformance was more pronounced insofar as α was lower.

Finally, the worst performers were *Dominance* and *Maximin*. On the graphical comparisons, *Dominance* sometimes did slightly better than *Maximin*, but the differences were marginal. These two players performed badly in comparison to other players, regardless of the bias. However, they performed better insofar as the bias was extreme and insofar as α was higher. These latter tendencies are especially troubling for these rules, because they are essentially the factors that can accelerate a decision rule's convergence towards expected payoff maximisation. Ceteris paribus, as the coin bias becomes more extreme and α is higher, players make more similar decisions. Therefore, *Dominance* and *Maximin* perform badly exactly when they behave differently from other players.

Why did *Dominance* perform as badly as *Maximin* in the graphical comparisons, but better than *Maximin* in the aggregate comparisons? The graphical comparisons detect that both players have problems with their average performances. However, the aggregate comparisons provide the additional information that *Maximin* sometimes chooses a directly, whereas *Dominance* only chooses a by randomising, and thus the latter's selection benefits from the discounting of random choices in the λ measure. This discounting is to reflect the fact that *Dominance*'s choice is a product of "bad luck". This result demonstrates the usefulness of including both of these methods of comparison: it enables us to not just detect results, but also to identify the causes of these results in particular features of players' decision rules.

4.3. Comparisons with Stan

We begin with the information in **Tables 10 to 13**. According to our aggregate comparisons, there were no statistically significant differences between *Stan* and the *Leaders*. However, the averages were closer when $\alpha = 0.5$, and there is a theoretical explanation for this pattern: insofar as the significance level is higher, confidence interval estimation approximates conditionalization with a flat prior beta distribution. Finally, in addition to having identical overall performances in our tests, the *Leaders* (with $\alpha = 0.5$) and *Stan* each had no statistically significant differences in their performances across different biases. Hence, there were no statistically significant differences in their performances for particular biases. Yet, with $\alpha = 0.01$ or $\alpha = 0.05$, the performance of the *Leaders* relative to *Stan* became inconsistent across coin biases: except with a bias of 0.5, *Stan* gained a small but statistically significant advantage. These results were corroborated by our graphical comparisons.

Stan and the *Leaders* were both very consistent performers. Hence, the comparisons of the other players relative to *Stan* match those reported in Subsection 4.1, and for the same reasons. Even given $\alpha = 0.5$, the calibration players other than the *Leaders* did at least slightly worse than *Stan* under some circumstances. However, given $\alpha = 0.5$, these underperformances were not persistent. This is because, if a calibration player's significance level is high, then they rapidly obtain a fairly precise interval of estimates for the coin bias. All the decision rules that we investigated have the property that if the confidence interval is very narrow, then the rule approximates expected payoff maximisation.

5. Discussion

Our results are especially intriguing because calibration players possess the same background knowledge and data as *Stan*.[12] Overall, we found that calibration players could match *Stan*, but only with certain rules and $\alpha = 0.5$. The latter setting narrows the procedural differences between the calibration players and *Stan*, but the calibration inductive logics are still a genuinely distinct approach to updating and decisions. Thus, their ability to match *Stan* means that there is more than one way to perform well in our decision problem. In the following subsections, we discuss some implications and limitations of these results.

5.1. Significance levels

Confidence interval estimation in science normally involves much lower values of α than were optimal for the calibration players (regardless of their decision rule) in our decision problem, so our results may be puzzling. The explanation is that in Bernoulli trials even early samples generally provide some information about the bias of the coin. This information should not be overestimated, but $\alpha = 0.5$ still discounts the significance of small samples relative to just extrapolating (or, worse, generating a sticky credence from) early sample frequencies. In this respect, it is analogous to *Stan*'s prior, which slightly discounts early sample data via a small but non-zero factor (their flat prior) against rapid extrapolation. Hence, despite their deep differences, the equally strong performance of *Stan* and the top performing calibration players is explained by a common ability to use (but not overuse) evidence to make good decisions in the early part of tests.

Obviously, we do *not* argue that the strong results for the $\alpha = 0.5$ setting imply that this significance level should be used in all comparable decision problems. For example, in scientific research, the incentives that scientists have to report statistically significant results, along with the incentives of journal editors to publish such results rather than replication studies, create the problem of publication bias in favour of significance. This publication bias problem complicates and potentially invalidates inferences made by aggregating these studies' results [12,13]. For this reason, some have argued that *tougher* significance levels should be used in statistical research, at least in some contexts [16,23,10,1]. Our results are consistent with these proposals, since we are looking at one class of decisions: betting on events in a single-agent context. By contrast, in contexts such as science, where there are distorting factors in favour of publishing significant results, like publication bias, political bias, and funding bias, it is plausible that lower significance levels are appropriate.

However, our results do suggest that it might be advantageous for scientists to provide a wider range of results, with estimates (and perhaps even discussion of their wider implications) at a broad range of different significance levels. People use scientific results to make many different types of decisions; a significance level that is appropriate for theoretical reasoning may be too cautious for decision problems featuring pecuniary or political decisions. (It depends on our loss functions, and therefore our aims and values.) According to some philosophers of science, such as Gregor Betz, showing the sensitivity of results and estimates to different standards for uncertainty would also increase the objectivity of science, since the research findings would be less closely tied to the partly subjective choice of a level of uncertainty (or "inductive risk") that a scientist might make, such as choosing a significance level [2]. Our results do not determine the outcomes of these philosophical debates, but do inform them, by illustrating how different purposes can require different significance levels.

5.2. Common factors

In this subsection, we shall describe the respective common features of the most successful calibration players and less successful players. The former, the *Leaders*, never randomised. Instead, they made what they regarded as their "best bet" given the available information. This distinguished the *Leaders* from *Dominance*. Meanwhile, they were distinguished from most of the less successful players by not abstaining, a behaviour they shared with *MaxEnt*. Additionally, the *Leaders* had a decision rule that performed as well as any other, regardless of the coin bias, unlike *MaxEnt*.

The middling players (in terms of our aggregate comparisons) were *Dominance*, *Regret*, *MaxEnt*, and *Pessimist*. With the exception of *MaxEnt*, their underperformance was partly due to failing to adequately utilise the information in early samples of Bernoulli trials, such that they randomised and abstained in games where the *Leaders* did not. The possibility of abstaining from betting is sometimes cited as a strength of imprecise probability decision rules [7] but in this decision problem it misses some possible systematic profit opportunities. *MaxEnt* undervalues the information from early samples, but in a

[12] We thank an anonymous referee for emphasising this point.

different way: like *Stan*, *MaxEnt* begins with an equivocal prior, but revises it more slowly. Surprisingly, at least by 250 games, *MaxEnt* gained no advantages over the *Leaders* when the coin bias is 0.5.

The least successful performers, *Maximin* and *ORO*, underperform partly because they randomise and abstain the most. Note that these players are still successful in the sense that they avoid making losses, even under the most challenging (for statistical inference) setting of a 0.5 coin bias. Our aggregate comparisons show that they also rapidly approximate the distribution of the choices that are appropriate given the knowledge of the true coin bias. Thus, their underperformance is relative to other players. The problem is that, even if they want to keep their calibration inductive logic rather than adopt a traditional Bayesian update and decision procedure like *Stan*'s, these players could do better in our decision problem if they adopted a different decision rule, like E-Admissibility or the Hurwicz criterion with a suitable value of α.

5.3. Limitations and generality

How far do our results generalise? As noted above, the key factor in the differences among players is that they exploit some fundamental features of Bernoulli trials. A very wide range of events can be understood as approximately binomial and exchangeable, especially in the limit, as shown by the central limit theorem and its extensions.[13] Thus, while most decision problems do not feature strict Bernoulli trials, there are many problems that are partly or wholly approximations of Bernoulli trials problems. Below, we discuss some additional considerations for determining the generality of our results.

5.3.1. Interaction

One might worry about making such comparative assessments when players do not take other players' behaviours or rules into account when making their decisions. Is it legitimate to compare players, if these players are not trying to defeat each other in a particular test, but simply maximise their own (accounting) profits? Additionally, our players do not contemplate switching rules in a particular test. So, is there a sense in which the performance of other rules are really opportunity costs for them? However, evaluating players in terms of opportunity costs does not require modelling interactive decision-making. We can interpret the opportunity costs as the costs that are considered by someone choosing which rule to use. The importance of opportunity cost for rationality does not depend on interactive behaviour.

Of course, interactive behaviour can illustrate points such as the general importance of opportunity costs. Competing business, scientific research teams, and other decision-makers usually must perform well in terms of opportunity costs to win competitions. In sum, it makes sense to compare players in terms of opportunity costs, even though the tests are not interactive.

5.3.2. Complexity

One concern about our study's generality could be that rapid learning and use of information is beneficial in our decision problem, whereas imprecise probability decision rules are often proposed to systematise cautious choice-making under circumstances where rapid learning can be costly. Yet it proves to be surprisingly difficult to specify circumstances where using one's information is not useful, except in *ad hoc* ways. As a source of relevant reasons for using "cautious" decision rules, we shall examine some arguments discussed by Peter Walley in his influential book [60, 215].[14] In the rest of this subsection we shall argue that, while Walley's points reveal some limitations of our results, the overall generality is considerable.

First, Walley suggests that precise probability approaches might be inappropriate given "complex or unstable" physical or social processes. One might think that this concern justifies lower significance levels (i.e. lower α values) or more cautious decision rules. Yet rapid learning can still be beneficial when making decisions about such processes. For example, if one is making decisions about a multidimensional set of Bernoulli (or approximately Bernoulli) parameters and one has a meagre data stream with respect to any particular dimension, then being able to learn rapidly from small samples can be very beneficial.

But what about when the early samples include misleading information? If this information is more misleading than would occur simply from random sampling, then exchangeability does not hold: the probability of certain outcomes in early samples must be higher or lower than the long-run probabilities. Like earlier studies using agent-based modelling to test calibrated inductive logic players ([35,50] we have focused on a decision problem featuring exchangeable events. The extension of our tests to games with non-exchangeable events will vary with the details of the probability distributions. While exchangeability is a very common assumption (even when it is not strictly true) in statistics and decision theory, extending our study to problems without this assumption is an exciting but challenging topic for future research.

For an example of these challenges, note that if calibration players know that their early samples are misleading, then they can lower their value of α. Thus, players with rules that might be regarded as less "cautious", such as the *Leaders* or *Dominance*, can actually reason in a cautious way, in that they will discount information from early samples. Both Objective Bayesians ([17]) and Evidential Probabilists ([29, Chapter 15]) regard the value of α as something that can be adjusted

[13] However, recall the role of the short-run: *in the limit*, a problem with a binomial distribution may be equivalent, for practical purposes, to a problem with a normal distribution or another type of probability distribution, but not the short-run. Therefore, one should not overinterpret the importance of asymptotic results like the central limit theorem for the generalisation of our findings in this study.

[14] We thank an anonymous referee for this suggestion.

in accordance with contextual payoffs and background information. Therefore, while there are complex decision problems where the ranking of players' performances will differ from our article, the extent of such problems is itself complex.

5.3.3. Unknowability

Walley mentions two more situations where imprecise probabilities could be useful: "the processes are physically indeterminate, governed by imprecise chances... or... the parameters of a complete model are not estimable from data". [60, p. 215] In either situation, agents cannot learn a true precise probability distribution. Are cautious imprecise probabilities or decision rules better in such situations? It will depend on the details of the decision problem. In our study, there is a true, precise, and knowable coin bias. Different results might be obtained when this assumption is changed. However, recall that even when discounting early sample information is advantageous, this can be obtained either by lowering α or by adopting a decision rule that makes minimal use of early sample information.

5.3.4. Overview

When extrapolating from our results to other problems, it is crucial to note that changes to the games in tests would modify player's behaviour. For example, we noted above that if players have reasons to discount the informational value of early samples, then they can modify their α values. An additional example is the following: if the base game described in **Fig. 1** had an increased guaranteed return from abstaining, this would increase payoffs for players who tend to abstain due to internal features of their decision rules. However, it would also affect the choice behaviour of players who do not abstain in the current version of the decision problem. Hence, one cannot infer that players who abstain more given the current base game will do better relative to other players when its payoff structure is modified to reward more cautious behaviour.

Furthermore, note that many of the decision rules we have discussed were originally designed for reasoning under pure ignorance, or even uncertainty about a decision-maker's information regarding the problem. Our results do not extend to these contexts, where learning prior to decisions is not possible. Consequently, they should not be interpreted as a criticism of the applicability of these decision rules in all situations. Despite these limitations, our results will extend to many traditional problems in inductive logic and dynamic decision theory, where rapid acquisition and use of information from samples is helpful for average payoff performances.

5.4. Characterising caution

In this study, we have included a wide range of decision rules. The purposes of these decision rules can conflict. For example, maximising payoffs can conflict with minimising worst-case expected loss. As an anonymous referee notes, this creates a challenge for evaluating these decision rules, since no single dimension of evaluation will include the goals of each decision theorist developing or using these rules.

To some extent, we have mitigated this issue by looking at two dimensions of assessment. Our aggregate comparison method is a test of an agent's ability to rapidly convergence to behaving as if they knew the true bias. In contrast, our graphical comparisons assess players with respect to their average payoff performances and their development over tests. Hence, the graphical comparisons enable the detection of subperiod differences, so a player who performs well overall but poorly in the very short-run will be identified. For instance, if a decision rule D_1 is more likely to make comparative losses in the first 20 games, whereas a different rule D_2 performs the same or worse in aggregate but is less likely to make those losses, then the graphical comparisons can detect this difference. As it happens, the *Leaders* did best with respect to both dimensions of assessment, but this was not inevitable.

Nonetheless, the article's results do not provide a self-contained evaluation of these rules performances in every possible respect. For instance, an evaluator needs to weigh *Maximin*'s performance in our measure and graphs given this rule's other decision-theoretic properties, such as with respect to minimising expected losses in particular games. Additionally, there might be other aspects of the tests that could be analysed in further research, such as the relative frequencies of absolute losses, the variance of payoffs, and so on. We look forward to examining such issues in future research.

There is an intriguing conceptual challenge raised by the divergent performances of decision rules when paired with these calibration inductive logics. Many of these rules are justified by their proponents as being "cautious" [63]. We do not dispute the significance of such formal results. For instance, if we interpret a coin bias of 0.5 as a "worst case" scenario for calibration inductive logics in our decision problem (because this bias has the highest long-run frequency of unrepresentative samples) then it is true that *MaxEnt* can match *Stan* and the *Leaders*. So, in the sense of "cautious" that means being prepared for worst case scenarios, *MaxEnt* is cautious.

Yet there are other ways to characterise "cautious" decision-making. *Maximin*'s decision rule is also intended to be cautious, but in the sense that it minimises expected worst-case losses. One might also think that avoiding opportunity costs could also be a criterion for making "cautious" decisions. There are two types of opportunity cost that are relevant. Firstly, there are opportunity costs in terms of an agent's own potential expected payoffs. The insensitivity of some rules, like Maximin, to this type of opportunity costs is one of the arguments for the Minimax Regret rule [51, p. 28]. Secondly, there is the opportunity cost of using one decision rule over another. Suppose that an agent is attracted to a rule because it satisfies some set of desiderata. Such a player might perform as well as any other rule, but there is a risk of doing worse. Our tests reveal the size of these risks for this decision problem. These risks could clearly be important in commercial,

policy, military, or other contexts. Minimising such risks could be characterised as "cautious". In that second sense, the *Leaders* are cautious: relative to other players, they minimise, on average, the risk of comparative underperformance.

Given these conceptual difficulties in characterising caution, we expect ongoing debates about imprecise probability update and decision rules that will be pertinent to the evaluation of our results. What are the senses in which a rule can be "cautious"? To what extent can they be combined in a single rule? How should different criteria be incorporated into the analysis of a rule's performance? We hope that the agent-based modelling methodology that we have expanded in this article, via the λ measure and other innovations, will help these debates in decision theory and inductive logic.

6. Conclusion

For some readers, it is perhaps surprising that some calibration players can match *Stan*. There have only been a few detailed comparisons of calibration inductive logics to more common approaches, such as conditionalization with standard Bayesianism [54,68,35], conditionalizing with sets of probability functions [54,56,48], or Dempster-Shafer belief functions [3]. Our results provide additional reasons for researchers in imprecise probability theory and inductive logic to study calibration approaches in more detail and to contrast their performance with alternative approaches in a wider range of problems.

Unlike almost all earlier studies (with the exceptions of [35] and [50]) we have proceeded by testing players via an agent-based model, rather than thought experiments or proofs of their general properties. We do not dispute the value of the latter approaches, but agent-based modelling enables fine-grained and quantitative comparisons that are otherwise impossible.

Furthermore, by focusing on opportunity costs, our results provide a new and very important perspective to some long-standing debates. It is one thing to know that a decision rule like Maximin is "cautious" in the sense of focusing on worst-case scenarios. However, if this caution is achieved at the price of a greatly inferior relative performance in the short-run, then it might be a price that users of calibrated inductive logics are unwilling to pay.

Advocates of these inductive logics should be encouraged that their systems can match the performances of *Stan* in this classic decision problem. Good decision-making is a common raison d'être for Bayesianism. While our results do not impact many decision-theoretic arguments for Bayesianism (such as Dutch Book Arguments) they do show that, in the context of Bernoulli trials and in the short-run, someone who favours Evidential Probability or Objective Bayesianism need not systematically underperform. However, this result is sensitive to the value of α. On the other hand, just as standard Bayesians can adjust their priors to different types of problems, so users of calibration inductive logics can adjust α and similar parameters to adapt to different types of decision problems. Thus, our results inform but do not settle the relevant debates between these approaches.

The common strong performance of *Stan* and the *Leaders*, despite their fundamental differences, echoes the early history of artificial intelligence. Researchers discovered that employing quantitative factors (such as probabilities, Dempster-Shafer belief functions, or certainty factors) in developing automated reasoning systems improved the performance of those systems, relative to purely qualitative rules. Yet the exact details of these quantitative factors or even their methodological justifications were less important than just adding finer-grained content to the system's reasoning [46, p. 487].[15]

There are many possible directions for future research. Firstly, there are other ways to update imprecise probabilities, such as Imprecise Bayesianism (where one conditionalizes using a set of probability functions) [7] and alpha-cut (where one "cuts" out probability functions from such a set if the function's prior probability for one's new evidence is below a certain level) [8]. Secondly, there are other popular priors in statistics that a Bayesian might use, such as Jeffreys priors, which would imply a setting of $B(0.5, 0.5)$ for the Bayesian player's beta prior [22]. Does such a standard Bayesian set a higher benchmark for the calibration players? Similarly, some probability theorists distinguish among the different probabilities that are compatible (according to an inductive logic) with a particular estimate of relative frequencies [47]. In the context of our investigations, this could mean that expectations in the imprecise probability interval that are further away from more natural values (such as those occurring more often in nature) are weighted more heavily in players' reasoning.[16] It would be interesting to see if such an approach can augment the performances of the calibration players in our tests. Thirdly, there are many ways in which we could modify the decision problem, as we discussed in the preceding section. Overall, this study has provided the first agent-based modelling tests of methods for making decisions with calibration inductive logics, and revealed the superiority of some methods for this type of decision problem.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Mantas Radzvilas reports that financial support was provided by German Research Foundation (project SP 279/21-1, project no. 420094936). William Peden reports that financial support was provided by Research Grants Council, University Grants Committee, project no. SRFS2122-3H01 of the Hong Kong SAR, China. Francesco De Pretis reports that financial support was provided by University of Modena and Reggio Emilia.

[15] We thank Jon Williamson for this suggestion.

[16] We thank an anonymous referee for this suggestion.

Data availability

Data will be made available on request.

Acknowledgements

We thank Daniele Tortoli (University of Modena and Reggio Emilia, Italy) for his very valuable support in accelerating computations in our research. The work described in this article was partly supported by a Senior Research Fellowship award from the Research Grants Council of the Hong Kong SAR, China ("Philosophy of Contemporary and Future Science", Project no. SRFS2122-3H01), German Research Foundation project SP 279/21-1 (project no. 420094936), and the University of Modena and Reggio Emilia, Italy. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of this article. The Authors are grateful to the University of Modena and Reggio Emilia for providing open access for this article.

Appendix A. Additional tables

Table 2
Statistics regarding the conditions met by Stan; Games = {250, 1000}.

Coin Bias	Condition	STAN	
		250 Games	1000 Games
0.1	K_1	26.252 ± 0.667	102.507 ± 0.689
	K_2	223.748 ± 0.667	897.493 ± 0.689
	K_3	0	0
0.3	K_1	74.844 ± 1.053	300.704 ± 1.033
	K_2	175.156 ± 1.053	699.296 ± 1.033
	K_3	0	0
0.5	K_1	125.344 ± 1.149	500.205 ± 1.115
	K_2	124.656 ± 1.149	499.795 ± 1.115
	K_3	0	0
0.7	K_1	175.120 ± 1.104	699.271 ± 1.066
	K_2	74.880 ± 1.104	300.729 ± 1.066
	K_3	0	0
0.9	K_1	223.940 ± 0.683	898.902 ± 0.705
	K_2	26.060 ± 0.683	101.098 ± 0.705
	K_3	0	0

Mean values and related standard errors multiplied by $z = 1.96$ (5% significance level) of the number of times a given condition was met by the pure action players. Statistics based on 1000 tests, each comprising 250 and 1000 games. Conditions' legend provided in **Table 1**.

Table 3
Statistics regarding the conditions met by MaxEnt; Games = {250, 1000}; $\alpha = \{0.01, 0.05, 0.5\}$.

Coin Bias	Condition	MAXENT					
		$\alpha = 0.01$		$\alpha = 0.05$		$\alpha = 0.5$	
		250 Games	1000 Games	250 Games	1000 Games	250 Games	1000 Games
0.1	K_1	38.147 ± 0.397	126.021 ± 0.752	35.117 ± 0.382	120.157 ± 0.735	29.003 ± 0.357	108.255 ± 0.715
	K_2	211.853 ± 0.397	873.979 ± 0.752	214.883 ± 0.382	879.843 ± 0.735	220.997 ± 0.357	891.745 ± 0.715
	K_3	0	0	0	0	0	0
0.3	K_1	91.077 ± 0.536	333.323 ± 1.062	87.522 ± 0.532	325.574 ± 1.055	79.783 ± 0.529	309.380 ± 1.042
	K_2	158.923 ± 0.536	666.677 ± 1.062	162.478 ± 0.532	674.426 ± 1.055	170.217 ± 0.529	690.620 ± 1.042
	K_3	0	0	0	0	0	0
0.5	K_1	124.927 ± 0.483	499.688 ± 0.930	124.938 ± 0.484	499.720 ± 0.932	124.897 ± 0.503	499.883 ± 0.978
	K_2	125.073 ± 0.483	500.312 ± 0.930	125.062 ± 0.484	500.280 ± 0.932	125.103 ± 0.503	500.117 ± 0.978
	K_3	0	0	0	0	0	0
0.7	K_1	159.103 ± 0.544	667.083 ± 1.079	162.461 ± 0.549	674.550 ± 1.084	170.015 ± 0.539	690.613 ± 1.072
	K_2	90.897 ± 0.544	332.917 ± 1.079	87.539 ± 0.549	325.450 ± 1.084	79.985 ± 0.539	309.387 ± 1.072
	K_3	0	0	0	0	0	0
0.9	K_1	212.039 ± 0.410	875.672 ± 0.769	215.088 ± 0.396	881.477 ± 0.752	221.207 ± 0.363	893.265 ± 0.719
	K_2	37.961 ± 0.410	124.328 ± 0.769	34.912 ± 0.396	118.523 ± 0.752	28.793 ± 0.363	106.735 ± 0.719
	K_3	0	0	0	0	0	0

Mean values and related standard errors multiplied by $z = 1.96$ (5% significance level) of the number of times a given condition was met by the pure action players. Statistics based on 1000 tests, each comprising 250 and 1000 games. Conditions' legend provided in **Table 1**.

Table 4
Statistics regarding the conditions met by other calibration players; Games = 250; $\alpha = 0.01$.

Coin Bias	Condition	LEADERS	PESSIMIST	DOMINANCE	MAXIMIN	REGRET	ORO
0.1	K_1	27.587 ± 0.342	21.980 ± 0.312	16.484 ± 0.281	16.484 ± 0.281	23.833 ± 0.318	18.211 ± 0.297
	K_2	222.413 ± 0.342	216.926 ± 0.369	211.474 ± 0.399	211.474 ± 0.399	218.743 ± 0.362	213.222 ± 0.387
	K_3	0	11.094 ± 0.200	0	22.042 ± 0.272	7.424 ± 0.167	18.567 ± 0.255
	K_4	0	0	0	0	0	0
	K_5	0	0	0	0	0	0
	K_6	0	0	0	0	0	0
	K_7	0	0	22.042 ± 0.272	0	0	0
0.3	K_1	76.207 ± 0.518	68.086 ± 0.489	60.280 ± 0.471	60.280 ± 0.471	70.890 ± 0.502	63.542 ± 0.478
	K_2	173.793 ± 0.518	165.822 ± 0.528	157.677 ± 0.544	157.677 ± 0.544	168.474 ± 0.526	160.917 ± 0.538
	K_3	0	16.092 ± 0.234	0	32.043 ± 0.321	10.636 ± 0.192	25.541 ± 0.298
	K_4	0	0	0	0	0	0
	K_5	0	0	0	0	0	0
	K_6	0	0	0	0	0	0
	K_7	0	0	32.043 ± 0.321	0	0	0
0.5	K_1	124.998 ± 0.567	116.040 ± 0.572	107.351 ± 0.570	107.351 ± 0.570	119.055 ± 0.572	111.098 ± 0.571
	K_2	125.002 ± 0.567	116.266 ± 0.561	107.501 ± 0.552	107.501 ± 0.552	119.188 ± 0.566	111.362 ± 0.558
	K_3	0	17.694 ± 0.251	0	35.148 ± 0.322	11.757 ± 0.205	27.540 ± 0.307
	K_4	0	0	0	0	0	0
	K_5	0	0	0	0	0	0
	K_6	0	0	0	0	0	0
	K_7	0	0	35.148 ± 0.322	0	0	0
0.7	K_1	173.526 ± 0.535	165.680 ± 0.551	157.766 ± 0.557	157.766 ± 0.557	168.273 ± 0.544	160.985 ± 0.550
	K_2	76.474 ± 0.535	68.344 ± 0.514	60.237 ± 0.495	60.237 ± 0.495	71.114 ± 0.521	63.599 ± 0.504
	K_3	0	15.976 ± 0.236	0	31.997 ± 0.312	10.613 ± 0.198	25.416 ± 0.285
	K_4	0	0	0	0	0	0
	K_5	0	0	0	0	0	0
	K_6	0	0	0	0	0	0
	K_7	0	0	31.997 ± 0.312	0	0	0
0.9	K_1	222.628 ± 0.350	217.123 ± 0.379	211.597 ± 0.411	211.597 ± 0.411	218.954 ± 0.371	213.369 ± 0.400
	K_2	27.372 ± 0.350	21.966 ± 0.313	16.630 ± 0.285	16.630 ± 0.285	23.805 ± 0.323	18.330 ± 0.299
	K_3	0	10.911 ± 0.201	0	21.773 ± 0.283	7.241 ± 0.167	18.301 ± 0.260
	K_4	0	0	0	0	0	0
	K_5	0	0	0	0	0	0
	K_6	0	0	0	0	0	0
	K_7	0	0	21.773 ± 0.283	0	0	0

Mean values and related standard errors multiplied by $z = 1.96$ (5% significance level) of the number of times a given condition was met by the mixed action players. Statistics based on 1000 tests, each one comprising 250 games. Conditions' legend provided in **Table 1**.

Table 5
Statistics regarding the conditions met by other calibration players; Games = 1000; $\alpha = 0.01$.

Coin Bias	Condition	LEADERS	PESSIMIST	DOMINANCE	MAXIMIN	REGRET	ORO
0.1	K_1	104.218 ± 0.691	93.112 ± 0.664	82.134 ± 0.628	82.134 ± 0.628	96.803 ± 0.667	84.232 ± 0.637
	K_2	895.782 ± 0.691	884.648 ± 0.721	873.600 ± 0.753	873.600 ± 0.753	888.332 ± 0.716	875.702 ± 0.744
	K_3	0	22.240 ± 0.285	0	44.266 ± 0.399	14.865 ± 0.234	40.066 ± 0.381
	K_4	0	0	0	0	0	0
	K_5	0	0	0	0	0	0
	K_6	0	0	0	0	0	0
	K_7	0	0	44.266 ± 0.399	0	0	0
0.3	K_1	301.567 ± 1.031	284.991 ± 1.007	268.884 ± 1.002	268.884 ± 1.002	290.715 ± 1.021	272.933 ± 1.005
	K_2	698.433 ± 1.031	682.091 ± 1.054	665.431 ± 1.069	665.431 ± 1.069	687.512 ± 1.046	669.498 ± 1.063
	K_3	0	32.918 ± 0.341	0	65.685 ± 0.471	21.773 ± 0.279	57.569 ± 0.451
	K_4	0	0	0	0	0	0
	K_5	0	0	0	0	0	0
	K_6	0	0	0	0	0	0
	K_7	0	0	65.685 ± 0.471	0	0	0
0.5	K_1	500.214 ± 1.111	482.235 ± 1.113	464.419 ± 1.104	464.419 ± 1.104	488.228 ± 1.111	469.110 ± 1.106
	K_2	499.786 ± 1.111	481.901 ± 1.117	464.168 ± 1.120	464.168 ± 1.120	487.876 ± 1.110	468.971 ± 1.126
	K_3	0	35.864 ± 0.367	0	71.413 ± 0.490	23.896 ± 0.302	61.919 ± 0.475
	K_4	0	0	0	0	0	0
	K_5	0	0	0	0	0	0
	K_6	0	0	0	0	0	0
	K_7	0	0	71.413 ± 0.490	0	0	0
0.7	K_1	698.418 ± 1.064	682.155 ± 1.081	665.746 ± 1.089	665.746 ± 1.089	687.466 ± 1.074	669.715 ± 1.080
	K_2	301.582 ± 1.064	285.056 ± 1.038	268.732 ± 1.024	268.732 ± 1.024	290.544 ± 1.048	272.899 ± 1.029
	K_3	0	32.789 ± 0.326	0	65.522 ± 0.470	21.990 ± 0.281	57.386 ± 0.446
	K_4	0	0	0	0	0	0
	K_5	0	0	0	0	0	0
	K_6	0	0	0	0	0	0
	K_7	0	0	65.522 ± 0.470	0	0	0
0.9	K_1	897.216 ± 0.707	886.162 ± 0.732	875.230 ± 0.770	875.230 ± 0.770	889.868 ± 0.725	877.319 ± 0.762
	K_2	102.784 ± 0.707	91.846 ± 0.667	80.963 ± 0.635	80.963 ± 0.635	95.466 ± 0.684	83.023 ± 0.647
	K_3	0	21.992 ± 0.284	0	43.807 ± 0.415	14.666 ± 0.235	39.658 ± 0.390
	K_4	0	0	0	0	0	0
	K_5	0	0	0	0	0	0
	K_6	0	0	0	0	0	0
	K_7	0	0	43.807 ± 0.415	0	0	0

Mean values and related standard errors multiplied by $z = 1.96$ (5% significance level) of the number of times a given condition was met by the mixed action players. Statistics based on 1000 tests, each one comprising 1000 games. Conditions' legend provided in **Table 1**.

Table 6
Statistics regarding the conditions met by other calibration players; Games = 250; $\alpha = 0.05$.

Coin Bias	Condition	LEADERS	PESSIMIST	DOMINANCE	MAXIMIN	REGRET	ORO
0.1	K_1	26.812 ± 0.340	22.433 ± 0.316	18.203 ± 0.291	18.203 ± 0.291	23.906 ± 0.322	19.273 ± 0.299
	K_2	223.188 ± 0.340	218.943 ± 0.363	214.681 ± 0.383	214.681 ± 0.383	220.338 ± 0.357	215.753 ± 0.377
	K_3	0	8.624 ± 0.181	0	17.116 ± 0.238	5.756 ± 0.147	14.974 ± 0.230
	K_4	0	0	0	0	0	0
	K_5	0	0	0	0	0	0
	K_6	0	0	0	0	0	0
	K_7	0	0	17.116 ± 0.238	0	0	0
0.3	K_1	75.796 ± 0.518	69.535 ± 0.495	63.433 ± 0.477	63.433 ± 0.477	71.688 ± 0.505	65.450 ± 0.485
	K_2	174.204 ± 0.518	168.001 ± 0.527	161.702 ± 0.537	161.702 ± 0.537	170.049 ± 0.527	163.824 ± 0.534
	K_3	0	12.464 ± 0.208	0	24.865 ± 0.283	8.263 ± 0.170	20.726 ± 0.265
	K_4	0	0	0	0	0	0
	K_5	0	0	0	0	0	0
	K_6	0	0	0	0	0	0
	K_7	0	0	24.865 ± 0.283	0	0	0
0.5	K_1	124.992 ± 0.569	118.042 ± 0.574	111.224 ± 0.571	111.224 ± 0.571	120.293 ± 0.572	113.679 ± 0.574
	K_2	125.008 ± 0.569	118.186 ± 0.566	111.475 ± 0.560	111.475 ± 0.560	120.532 ± 0.569	113.868 ± 0.562
	K_3	0	13.772 ± 0.219	0	27.301 ± 0.288	9.175 ± 0.180	22.453 ± 0.277
	K_4	0	0	0	0	0	0
	K_5	0	0	0	0	0	0
	K_6	0	0	0	0	0	0
	K_7	0	0	27.301 ± 0.288	0	0	0
0.7	K_1	173.922 ± 0.536	167.758 ± 0.547	161.593 ± 0.558	161.593 ± 0.558	169.870 ± 0.539	163.689 ± 0.552
	K_2	76.078 ± 0.536	69.806 ± 0.523	63.514 ± 0.506	63.514 ± 0.506	71.927 ± 0.522	65.619 ± 0.513
	K_3	0	12.436 ± 0.209	0	24.893 ± 0.284	8.203 ± 0.177	20.692 ± 0.269
	K_4	0	0	0	0	0	0
	K_5	0	0	0	0	0	0
	K_6	0	0	0	0	0	0
	K_7	0	0	24.893 ± 0.284	0	0	0
0.9	K_1	223.369 ± 0.346	219.157 ± 0.374	214.862 ± 0.397	214.862 ± 0.397	220.584 ± 0.365	215.936 ± 0.391
	K_2	26.631 ± 0.346	22.429 ± 0.317	18.298 ± 0.299	18.298 ± 0.299	23.844 ± 0.324	19.342 ± 0.306
	K_3	0	8.414 ± 0.177	0	16.840 ± 0.248	5.572 ± 0.147	14.722 ± 0.235
	K_4	0	0	0	0	0	0
	K_5	0	0	0	0	0	0
	K_6	0	0	0	0	0	0
	K_7	0	0	16.840 ± 0.248	0	0	0

Mean values and related standard errors multiplied by $z = 1.96$ (5% significance level) of the number of times a given condition was met by the mixed action players. Statistics based on 1000 tests, each one comprising 250 games. Conditions' legend provided in **Table 1**.

Table 7
Statistics regarding the conditions met by other calibration players; Games = 1000; $\alpha = 0.05$.

Coin Bias	Condition	LEADERS	PESSIMIST	DOMINANCE	MAXIMIN	REGRET	ORO
0.1	K_1	103.238 ± 0.690	94.634 ± 0.669	86.229 ± 0.643	86.229 ± 0.643	97.550 ± 0.674	87.489 ± 0.643
	K_2	896.762 ± 0.690	888.181 ± 0.719	879.641 ± 0.736	879.641 ± 0.736	891.004 ± 0.713	880.930 ± 0.732
	K_3	0	17.185 ± 0.255	0	34.130 ± 0.346	11.446 ± 0.211	31.581 ± 0.337
	K_4	0	0	0	0	0	0
	K_5	0	0	0	0	0	0
	K_6	0	0	0	0	0	0
	K_7	0	0	34.130 ± 0.346	0	0	0
0.3	K_1	301.069 ± 1.033	288.466 ± 1.015	275.866 ± 1.006	275.866 ± 1.006	292.665 ± 1.022	278.370 ± 1.011
	K_2	698.931 ± 1.033	686.358 ± 1.048	673.650 ± 1.061	673.650 ± 1.061	690.554 ± 1.040	676.216 ± 1.058
	K_3	0	25.176 ± 0.298	0	50.484 ± 0.418	16.781 ± 0.248	45.414 ± 0.400
	K_4	0	0	0	0	0	0
	K_5	0	0	0	0	0	0
	K_6	0	0	0	0	0	0
	K_7	0	0	50.484 ± 0.418	0	0	0
0.5	K_1	500.205 ± 1.113	486.378 ± 1.115	472.629 ± 1.110	472.629 ± 1.110	490.921 ± 1.115	475.625 ± 1.113
	K_2	499.795 ± 1.113	486.004 ± 1.112	472.386 ± 1.121	472.386 ± 1.121	490.616 ± 1.112	475.339 ± 1.121
	K_3	0	27.618 ± 0.324	0	54.985 ± 0.434	18.463 ± 0.263	49.036 ± 0.417
	K_4	0	0	0	0	0	0
	K_5	0	0	0	0	0	0
	K_6	0	0	0	0	0	0
	K_7	0	0	54.985 ± 0.434	0	0	0
0.7	K_1	698.920 ± 1.066	686.208 ± 1.077	673.682 ± 1.091	673.682 ± 1.091	690.556 ± 1.072	676.234 ± 1.088
	K_2	301.080 ± 1.066	288.365 ± 1.043	275.837 ± 1.029	275.837 ± 1.029	292.597 ± 1.050	278.391 ± 1.033
	K_3	0	25.427 ± 0.292	0	50.481 ± 0.404	16.847 ± 0.249	45.375 ± 0.389
	K_4	0	0	0	0	0	0
	K_5	0	0	0	0	0	0
	K_6	0	0	0	0	0	0
	K_7	0	0	50.481 ± 0.404	0	0	0
0.9	K_1	898.170 ± 0.704	889.752 ± 0.729	881.251 ± 0.753	881.251 ± 0.753	892.565 ± 0.718	882.491 ± 0.746
	K_2	101.830 ± 0.704	93.365 ± 0.675	85.041 ± 0.652	85.041 ± 0.652	96.173 ± 0.686	86.275 ± 0.657
	K_3	0	16.883 ± 0.252	0	33.708 ± 0.357	11.262 ± 0.204	31.234 ± 0.344
	K_4	0	0	0	0	0	0
	K_5	0	0	0	0	0	0
	K_6	0	0	0	0	0	0
	K_7	0	0	33.708 ± 0.357	0	0	0

Mean values and related standard errors multiplied by $z = 1.96$ (5% significance level) of the number of times a given condition was met by the mixed action players. Statistics based on 1000 tests, each one comprising 1000 games. Conditions' legend provided in **Table 1**.

Table 8
Statistics regarding the conditions met by other calibration players; Games = 250; $\alpha = 0.5$.

Coin Bias	Condition	LEADERS	PESSIMIST	DOMINANCE	MAXIMIN	REGRET	ORO
0.1	K_1	25.683 ± 0.336	23.987 ± 0.324	22.287 ± 0.316	22.287 ± 0.316	24.527 ± 0.330	22.513 ± 0.318
	K_2	224.317 ± 0.336	222.633 ± 0.346	220.969 ± 0.358	220.969 ± 0.358	223.171 ± 0.343	221.199 ± 0.356
	K_3	0	3.380 ± 0.113	0	6.744 ± 0.158	2.302 ± 0.091	6.288 ± 0.154
	K_4	0	0	0	0	0	0
	K_5	0	0	0	0	0	0
	K_6	0	0	0	0	0	0
	K_7	0	0	6.744 ± 0.158	0	0	0
0.3	K_1	75.247 ± 0.519	72.872 ± 0.514	70.471 ± 0.504	70.471 ± 0.504	73.671 ± 0.514	70.862 ± 0.504
	K_2	174.753 ± 0.519	172.380 ± 0.522	170.008 ± 0.531	170.008 ± 0.531	173.164 ± 0.520	170.402 ± 0.530
	K_3	0	4.748 ± 0.132	0	9.521 ± 0.180	3.165 ± 0.110	8.736 ± 0.175
	K_4	0	0	0	0	0	0
	K_5	0	0	0	0	0	0
	K_6	0	0	0	0	0	0
	K_7	0	0	9.521 ± 0.180	0	0	0
0.5	K_1	124.972 ± 0.572	122.247 ± 0.574	119.712 ± 0.574	119.712 ± 0.574	123.136 ± 0.575	120.136 ± 0.573
	K_2	125.028 ± 0.572	122.385 ± 0.570	119.904 ± 0.572	119.904 ± 0.572	123.269 ± 0.571	120.373 ± 0.572
	K_3	0	5.368 ± 0.133	0	10.384 ± 0.188	3.595 ± 0.111	9.491 ± 0.181
	K_4	0	0	0	0	0	0
	K_5	0	0	0	0	0	0
	K_6	0	0	0	0	0	0
	K_7	0	0	10.384 ± 0.188	0	0	0
0.7	K_1	174.486 ± 0.537	172.099 ± 0.541	169.797 ± 0.543	169.797 ± 0.543	172.908 ± 0.542	170.184 ± 0.542
	K_2	75.514 ± 0.537	73.163 ± 0.536	70.756 ± 0.527	70.756 ± 0.527	73.980 ± 0.533	71.138 ± 0.527
	K_3	0	4.738 ± 0.131	0	9.447 ± 0.185	3.112 ± 0.108	8.678 ± 0.179
	K_4	0	0	0	0	0	0
	K_5	0	0	0	0	0	0
	K_6	0	0	0	0	0	0
	K_7	0	0	9.447 ± 0.185	0	0	0
0.9	K_1	224.456 ± 0.342	222.808 ± 0.354	221.175 ± 0.363	221.175 ± 0.363	223.376 ± 0.351	221.407 ± 0.361
	K_2	25.544 ± 0.342	23.902 ± 0.334	22.315 ± 0.322	22.315 ± 0.322	24.468 ± 0.334	22.508 ± 0.323
	K_3	0	3.290 ± 0.111	0	6.510 ± 0.152	2.156 ± 0.092	6.085 ± 0.146
	K_4	0	0	0	0	0	0
	K_5	0	0	0	0	0	0
	K_6	0	0	0	0	0	0
	K_7	0	0	6.510 ± 0.152	0	0	0

Mean values and related standard errors multiplied by $z = 1.96$ (5% significance level) of the number of times a given condition was met by the mixed action players. Statistics based on 1000 tests, each one comprising 250 games. Conditions' legend provided in **Table 1**.

Table 9
Statistics regarding the conditions met by other calibration players; Games = 1000; $\alpha = 0.5$.

Coin Bias	Condition	LEADERS	PESSIMIST	DOMINANCE	MAXIMIN	REGRET	ORO
0.1	K_1	101.860 ± 0.689	98.676 ± 0.680	95.486 ± 0.671	95.486 ± 0.671	99.712 ± 0.684	95.737 ± 0.672
	K_2	898.140 ± 0.689	894.888 ± 0.699	891.717 ± 0.715	891.717 ± 0.715	895.971 ± 0.696	891.963 ± 0.713
	K_3	0	6.436 ± 0.158	0	12.797 ± 0.220	4.317 ± 0.127	12.300 ± 0.217
	K_4	0	0	0	0	0	0
	K_5	0	0	0	0	0	0
	K_6	0	0	0	0	0	0
	K_7	0	0	12.797 ± 0.220	0	0	0
0.3	K_1	300.388 ± 1.035	295.743 ± 1.028	291.105 ± 1.026	291.105 ± 1.026	297.280 ± 1.031	291.539 ± 1.025
	K_2	699.612 ± 1.035	695.106 ± 1.037	690.411 ± 1.043	690.411 ± 1.043	696.589 ± 1.036	690.856 ± 1.042
	K_3	0	9.151 ± 0.184	0	18.484 ± 0.257	6.131 ± 0.153	17.605 ± 0.253
	K_4	0	0	0	0	0	0
	K_5	0	0	0	0	0	0
	K_6	0	0	0	0	0	0
	K_7	0	0	18.484 ± 0.257	0	0	0
0.5	K_1	500.186 ± 1.117	495.062 ± 1.119	490.092 ± 1.115	490.092 ± 1.115	496.686 ± 1.118	490.583 ± 1.117
	K_2	499.814 ± 1.117	494.675 ± 1.116	489.744 ± 1.114	489.744 ± 1.114	496.370 ± 1.116	490.275 ± 1.115
	K_3	0	10.263 ± 0.192	0	20.164 ± 0.273	6.944 ± 0.158	19.142 ± 0.268
	K_4	0	0	0	0	0	0
	K_5	0	0	0	0	0	0
	K_6	0	0	0	0	0	0
	K_7	0	0	20.164 ± 0.273	0	0	0
0.7	K_1	699.592 ± 1.067	694.965 ± 1.074	690.395 ± 1.074	690.395 ± 1.074	696.512 ± 1.071	690.831 ± 1.074
	K_2	300.408 ± 1.067	295.786 ± 1.059	291.096 ± 1.049	291.096 ± 1.049	297.335 ± 1.057	291.533 ± 1.050
	K_3	0	9.249 ± 0.185	0	18.509 ± 0.255	6.153 ± 0.151	17.636 ± 0.251
	K_4	0	0	0	0	0	0
	K_5	0	0	0	0	0	0
	K_6	0	0	0	0	0	0
	K_7	0	0	18.509 ± 0.255	0	0	0
0.9	K_1	899.501 ± 0.704	896.380 ± 0.713	893.233 ± 0.719	893.233 ± 0.719	897.462 ± 0.710	893.493 ± 0.718
	K_2	100.499 ± 0.704	97.308 ± 0.694	94.190 ± 0.683	94.190 ± 0.683	98.437 ± 0.696	94.410 ± 0.684
	K_3	0	6.312 ± 0.156	0	12.577 ± 0.215	4.101 ± 0.128	12.097 ± 0.210
	K_4	0	0	0	0	0	0
	K_5	0	0	0	0	0	0
	K_6	0	0	0	0	0	0
	K_7	0	0	12.577 ± 0.215	0	0	0

Mean values and related standard errors multiplied by $z = 1.96$ (5% significance level) of the number of times a given condition was met by the mixed action players. Statistics based on 1000 tests, each one comprising 1000 games. Conditions' legend provided in **Table 1**.

Table 10
Aggregate comparisons: Stan.

Coin Bias	Games	STAN
0.1	250	99.738 (0; 99.477)
	1000	99.861 (0; 99.721)
0.3	250	99.861 (0; 99.723)
	1000	99.950 (0; 99.899)
0.5	250	99.996 (0; 99.993)
	1000	99.980 (0; 99.959)
0.7	250	99.779 (0; 99.558)
	1000	99.948 (0; 99.896)
0.9	250	99.776 (0; 99.552)
	1000	99.939 (0; 99.878)
Average Performance	250	99.830 ± 0.097
	1000	99.935 ± 0.039

Performance measurements for 250 games and 1000 games, for each coin bias, with abstain and goodness of fit parameter values Λ_1, Λ_2 reported between brackets. At the bottom, average performance with confidence intervals shown at 5% significance level.

Table 11
Aggregate comparisons: Calibration players $\alpha = 0.01$.

Coin Bias	Games	LEADERS	MAXENT	DOMINANCE	REGRET	PESSIMIST	ORO	MAXIMIN
0.1	250	99.423 (0; 98.850)	97.034 (0; 94.157)	98.153 (02.939; 99.258)	98.136 (02.970; 99.255)	96.987 (04.438; 98.433)	94.644 (07.427; 96.761)	93.557 (08.817; 95.992)
	1000	99.765 (0; 99.531)	98.544 (0; 97.109)	99.041 (01.476; 99.560)	99.031 (01.487; 99.551)	98.456 (2.224; 99.141)	97.067 (04.007; 98.153)	96.740 (04.427; 97.920)
0.3	250	99.655 (0; 99.310)	95.296 (0; 90.813)	96.484 (04.272; 97.246)	96.489 (04.254; 97.237)	94.606 (06.437; 95.661)	91.379 (10.216; 93.003)	89.141 (12.817; 91.142)
	1000	99.888 (0; 99.776)	97.591 (0; 95.240)	98.172 (02.189; 98.535)	98.179 (02.177; 98.536)	97.219 (03.292; 97.732)	95.111 (05.757; 95.986)	94.414 (06.568; 95.407)
0.5	250	99.999 (0; 99.998)	99.971 (0; 99.942)	95.314 (04.686; 95.314)	95.297 (04.703; 95.297)	92.922 (07.078; 92.922)	88.984 (11.016; 88.984)	85.941 (14.059; 85.941)
	1000	99.979 (0; 99.957)	99.969 (0; 99.938)	97.620 (02.380; 97.620)	97.610 (02.390; 97.610)	96.414 (03.586; 96.414)	93.808 (06.192; 93.808)	92.859 (07.141; 92.859)
0.7	250	99.578 (0; 99.158)	95.350 (0; 90.916)	96.481 (04.266; 97.235)	96.521 (04.245; 97.292)	94.669 (06.390; 95.740)	91.421 (10.166; 93.037)	89.149 (12.799; 91.140)
	1000	99.887 (0; 99.774)	97.620 (0; 95.298)	98.170 (02.184; 98.525)	98.158 (02.199; 98.517)	97.230 (03.279; 97.741)	95.123 (05.739; 95.992)	94.421 (06.552; 95.405)
0.9	250	99.471 (0; 98.946)	97.077 (0; 94.240)	98.183 (02.903; 99.282)	98.173 (02.896; 99.253)	97.026 (04.364; 98.436)	94.725 (07.320; 96.815)	93.643 (08.709; 96.056)
	1000	99.845 (0; 99.691)	98.639 (0; 97.297)	98.991 (01.460; 99.444)	98.983 (01.467; 99.434)	98.414 (02.199; 99.031)	97.037 (03.966; 98.050)	96.714 (04.381; 97.821)
Average Performance	250	99.625 ± 0.200	96.946 ± 1.666	96.923 ± 1.080	96.923 ± 1.076	95.242 ± 1.540	92.231 ± 2.146	90.286 ± 2.889
	1000	99.873 ± 0.068	98.473 ± 0.852	98.399 ± 0.532	98.392 ± 0.531	97.547 ± 0.768	95.629 ± 1.231	95.030 ± 1.468

Performance measurements for 250 games and 1000 games, with Λ_1, Λ_2 values reported between brackets. The players are ordered by their average performances. Confidence intervals at 5% significance level.

Table 12
Aggregate comparisons: Calibration players $\alpha = 0.05$.

Coin Bias	Games	LEADERS	MAXENT	DOMINANCE	REGRET	PESSIMIST	ORO	MAXIMIN
0.1	250	99.597	97.726	98.535	98.523	97.621	95.640	94.970
		(0; 99.195)	(0; 95.504)	(03.315; 99.358)	(02.302; 99.355)	(03.450; 98.704)	(05.990; 97.298)	(06.846; 96.823)
	1000	99.820	98.874	99.261	99.254	98.805	97.686	97.488
		(0; 99.640)	(0; 97.760)	(01.138; 99.661)	(01.145; 99.655)	(01.719; 99.332)	(03.158; 98.537)	(3.413; 98.397)
0.3	250	99.772	96.356	97.256	97.259	95.811	92.978	91.561
		(0; 99.545)	(0; 92.845)	(03.315; 97.830)	(03.305; 97.826)	(04.986; 96.614)	(08.290; 94.264)	(09.946; 93.094)
	1000	99.924	98.156	98.589	98.591	97.871	96.136	95.701
		(0; 99.847)	(0; 96.347)	(01.683; 98.861)	(01.678; 98.862)	(02.518; 98.262)	(04.541; 96.818)	(05.048; 96.457)
0.5	250	99.997	99.975	96.360	96.330	94.491	91.019	89.080
		(0; 99.994)	(0; 99.950)	(03.640; 96.360)	(03.670; 96.330)	(05.509; 94.491)	(08.891; 91.019)	(10.920; 89.080)
	1000	99.979	99.972	98.167	98.154	97.238	95.096	94.502
		(0; 99.959)	(0; 99.944)	(01.833; 98.167)	(01.846; 98.154)	(02.762; 97.238)	(04.904; 95.096)	(05.498; 94.502)
0.7	250	99.692	96.351	97.263	97.303	95.850	93.007	91.562
		(0; 99.384)	(0; 92.835)	(03.319; 97.849)	(03.281; 97.891)	(04.974; 96.681)	(08.277; 94.309)	(09.957; 93.107)
	1000	99.923	98.165	98.588	98.585	97.851	96.139	95.701
		(0; 99.846)	(0; 96.364)	(01.683; 98.860)	(01.685; 98.855)	(02.543; 98.245)	(04.537; 96.821)	(05.048; 96.456)
0.9	250	99.637	97.773	98.556	98.553	97.668	95.709	95.049
		(0; 99.275)	(0; 95.595)	(02.245; 99.363)	(02.229; 99.341)	(03.366; 98.712)	(05.889; 97.334)	(06.736; 96.869)
	1000	99.898	98.966	99.210	99.204	98.766	97.652	97.459
		(0; 99.797)	(0; 97.942)	(01.124; 99.544)	(01.126; 99.535)	(01.688; 99.223)	(03.123; 98.433)	(03.371; 98.296)
Average	250	99.739 ± 0.139	97.636 ± 1.299	97.594 ± 0.827	97.594 ± 0.829	96.288 ± 1.187	93.671 ± 1.752	92.444 ± 2.236
Performance	1000	99.909 ± 0.051	98.827 ± 0.653	98.763 ± 0.407	98.758 ± 0.408	98.106 ± 0.588	96.542 ± 0.976	96.170 ± 1.128

Performance measurements for 250 games and 1000 games, with Λ_1, Λ_2 values reported between brackets. The players are ordered by their average performances. Confidence intervals at 5% significance level.

Table 13
Aggregate comparisons: Calibration players $\alpha = 0.5$.

Coin Bias	Games	LEADERS	MAXENT	DOMINANCE	REGRET	PESSIMIST	ORO	MAXIMIN
0.1	250	99.848	99.106	99.417	99.404	99.068	98.158	98.016
		(0; 99.696)	(0; 98.221)	(00.899; 99.735)	(00.921; 99.730)	(01.352; 99.490)	(02.515; 98.836)	(02.698; 98.736)
	1000	99.897	99.540	99.752	99.747	99.583	99.127	99.088
		(0; 99.793)	(0; 99.083)	(00.427; 99.931)	(00.432; 99.926)	(00.644; 99.811)	(01.230; 99.484)	(01.280; 99.457)
0.3	250	99.929	98.624	98.938	98.944	98.400	97.030	96.760
		(0; 99.859)	(0; 97.267)	(01.269; 99.146)	(01.266; 99.154)	(01.899; 98.700)	(03.494; 97.557)	(03.808; 97.333)
	1000	99.972	99.328	99.482	99.484	99.225	98.500	98.425
		(0; 99.945)	(0; 98.660)	(00.616; 99.580)	(00.613; 99.582)	(00.915; 99.365)	(01.760; 98.762)	(01.848; 98.700)
0.5	250	99.989	99.959	98.615	98.562	97.853	96.204	95.846
		(0; 99.978)	(0; 99.918)	(01.385; 98.615)	(01.438; 98.562)	(02.147; 97.853)	(03.796; 96.204)	(04.154; 95.846)
	1000	99.981	99.988	99.328	99.306	98.974	98.086	97.984
		(0; 99.963)	(0; 99.977)	(00.672; 99.328)	(00.694; 99.306)	(01.026; 98.974)	(01.914; 98.086)	(02.016; 97.984)
0.7	250	99.853	98.565	98.975	98.994	98.436	97.078	96.814
		(0; 99.706)	(0; 97.151)	(01.260; 99.210)	(01.245; 99.233)	(01.895; 98.768)	(03.471; 97.630)	(03.779; 97.410)
	1000	99.971	99.327	99.481	99.484	99.219	98.498	98.423
		(0; 99.942)	(0; 98.659)	(00.617; 99.579)	(00.615; 99.584)	(00.925; 99.363)	(01.764; 98.760)	(01.851; 98.698)
0.9	250	99.879	99.154	99.426	99.426	99.073	98.203	98.074
		(0; 99.758)	(0; 98.314)	(00.868; 99.720)	(00.862; 99.715)	(01.316; 99.463)	(02.434; 98.844)	(02.604; 98.756)
	1000	99.972	99.625	99.695	99.703	99.530	99.079	99.043
		(0; 99.945)	(0; 99.252)	(00.419; 99.810)	(00.410; 99.816)	(00.631; 99.691)	(01.210; 99.369)	(01.258; 99.344)
Average	250	99.900 ± 0.052	99.082 ± 0.490	99.074 ± 0.304	99.066 ± 0.315	98.566 ± 0.452	97.335 ± 0.742	97.102 ± 0.827
Performance	1000	99.959 ± 0.030	99.562 ± 0.238	99.548 ± 0.152	99.545 ± 0.158	99.306 ± 0.220	98.658 ± 0.386	98.593 ± 0.410

Performance measurements for 250 games and 1000 games, with Λ_1, Λ_2 values reported between brackets. The players are ordered by their average performances. Confidence intervals at 5% significance level.

Appendix B. Additional figures

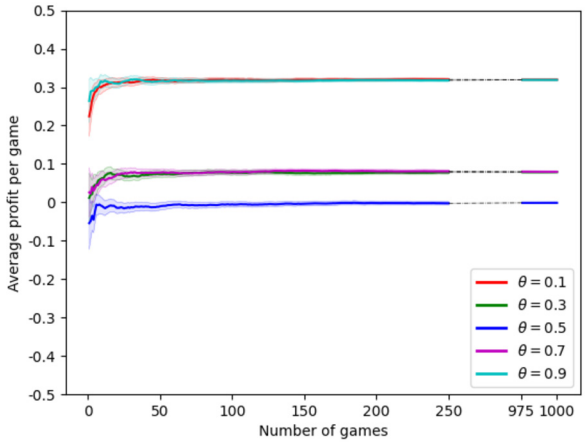
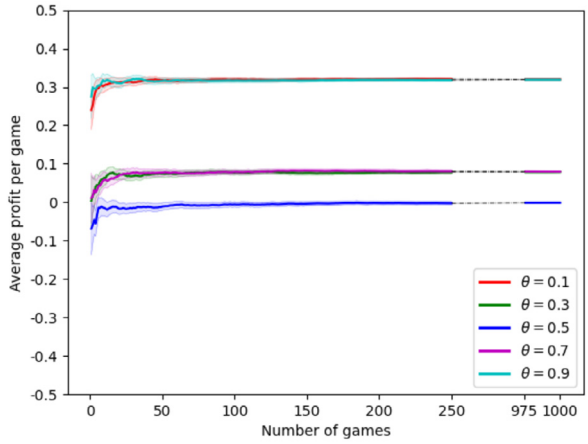
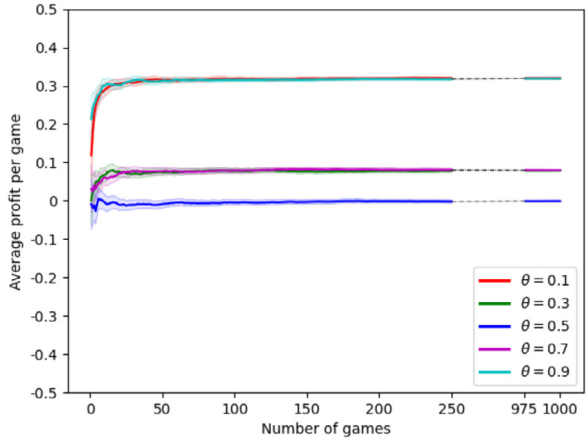


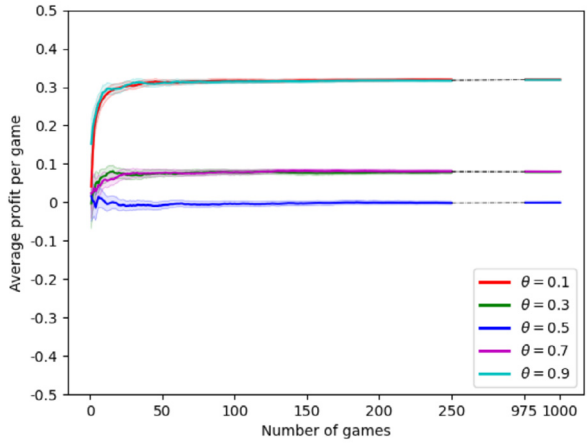
Fig. 2. Stan. Average profit per bet (Y-axis) against number of bets (X-axis) Coloured lines are the averages. The confidence intervals around them are calculated at the 0.95 level. (For interpretation of the colours in the figure(s), the reader is referred to the web version of this article.)



(a) $\alpha = 0.5$

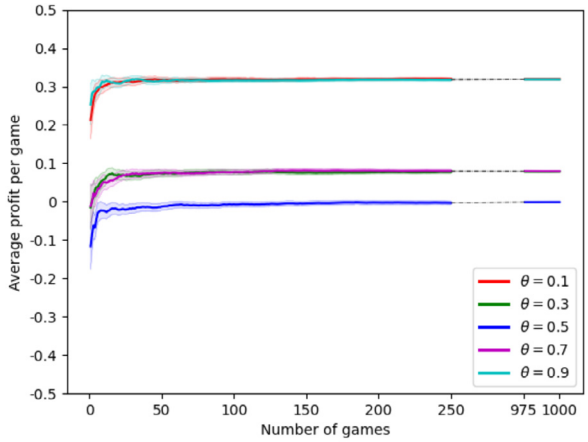


(b) $\alpha = 0.05$

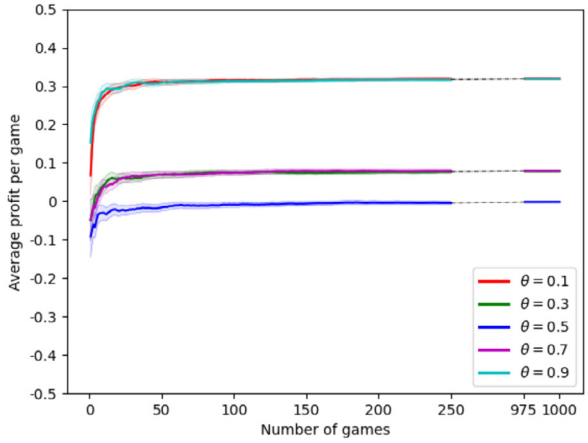


(c) $\alpha = 0.01$

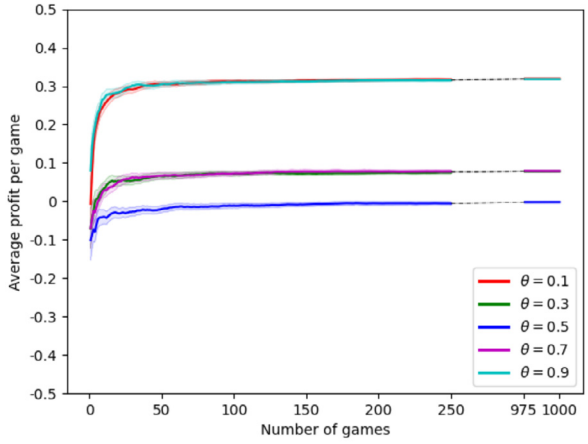
Fig. 3. Leaders. Average profit per bet (Y-axis) against number of bets (X-axis) In each subcaption, we report values for α. Coloured lines are the averages. The confidence intervals around them are calculated at the 0.95 level.



(a) $\alpha = 0.5$

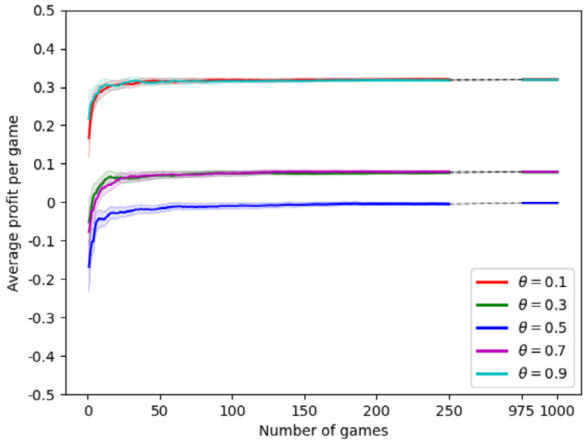


(b) $\alpha = 0.05$

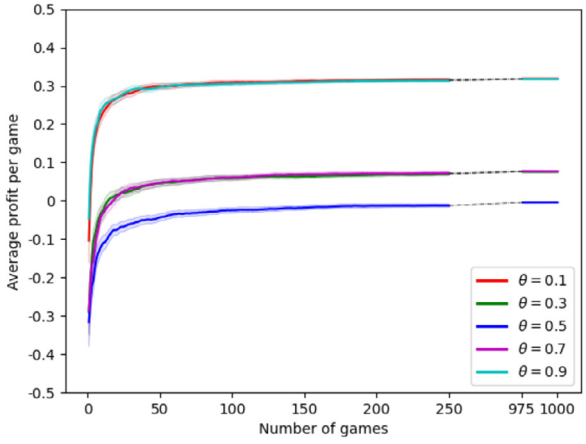


(c) $\alpha = 0.01$

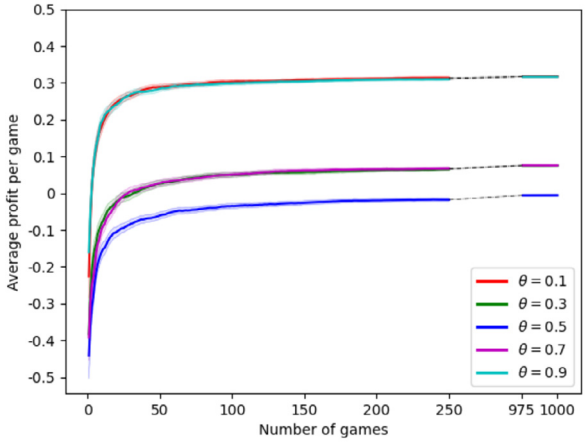
Fig. 4. Pessimist. Average profit per bet (Y-axis) against number of bets (X-axis) In each subcaption, we report values for α. Coloured lines are the averages. The confidence intervals around them are calculated at the 0.95 level.



(a) $\alpha = 0.5$

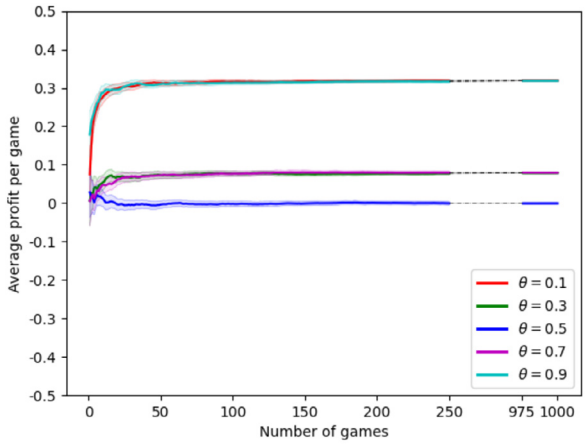


(b) $\alpha = 0.05$

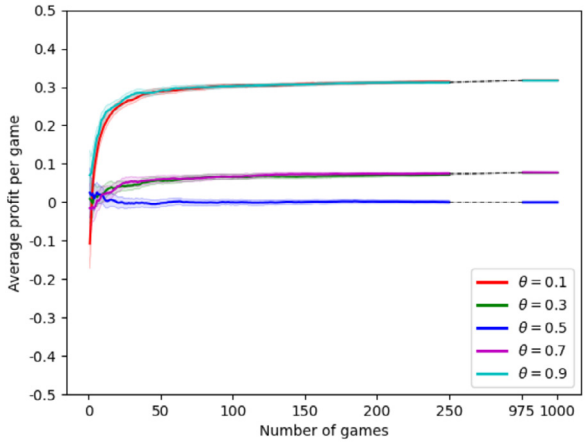


(c) $\alpha = 0.01$

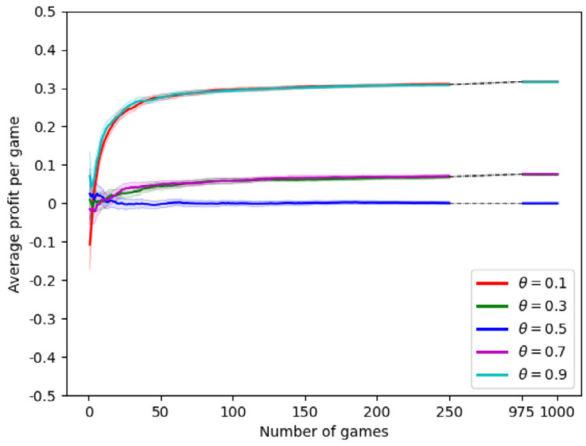
Fig. 5. Dominance. Average profit per bet (Y-axis) against number of bets (X-axis) In each subcaption, we report values for α. Coloured lines are the averages. The confidence intervals around them are calculated at the 0.95 level.



(a) $\alpha = 0.5$

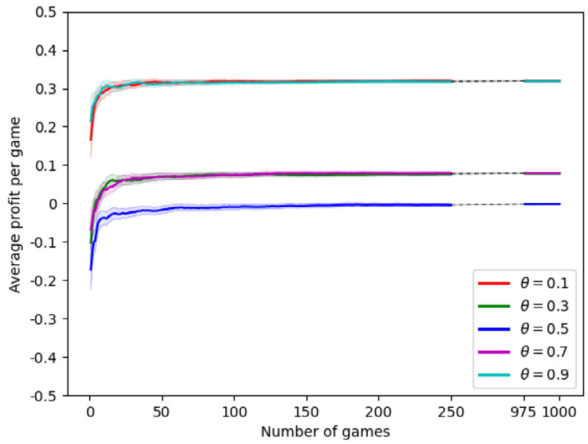


(b) $\alpha = 0.05$

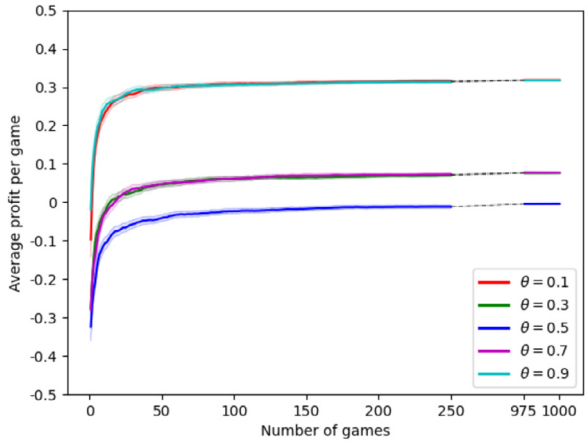


(c) $\alpha = 0.01$

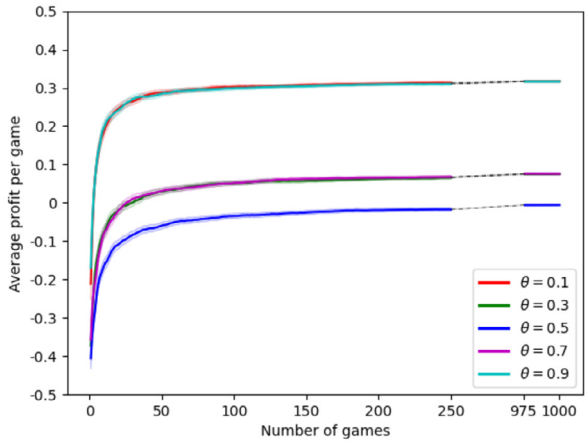
Fig. 6. MaxEnt. Average profit per bet (Y-axis) against number of bets (X-axis) In each subcaption, we report values for α. Coloured lines are the averages. The confidence intervals around them are calculated at the 0.95 level.



(a) $\alpha = 0.5$

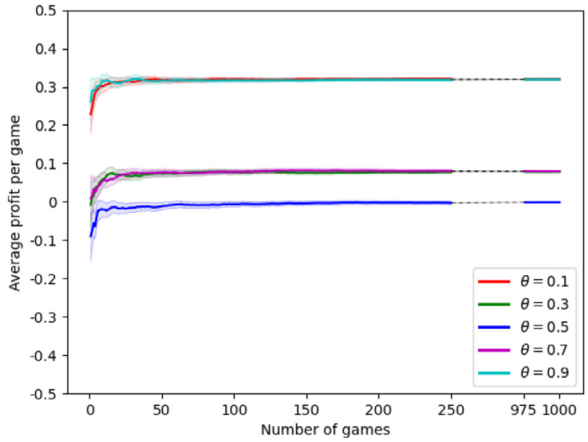


(b) $\alpha = 0.05$

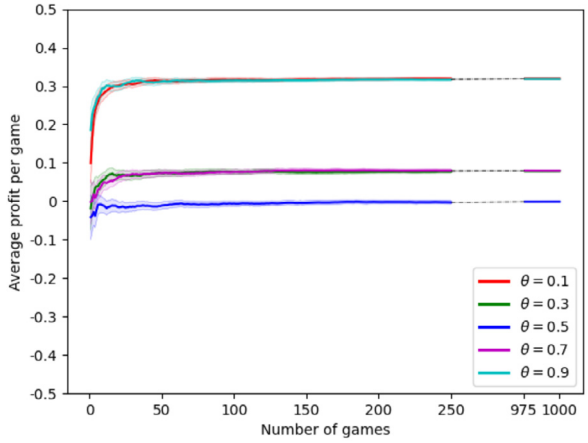


(c) $\alpha = 0.01$

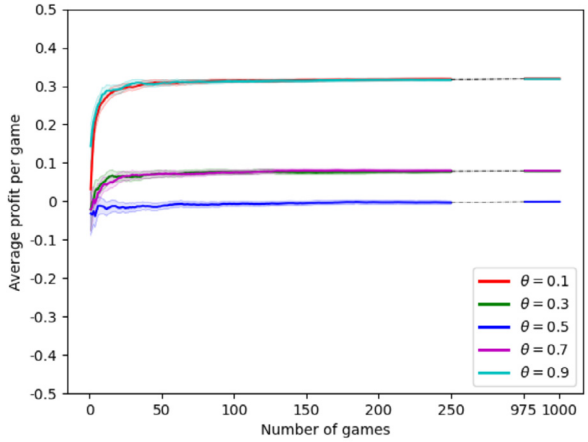
Fig. 7. Maximin. Average profit per bet (Y-axis) against number of bets (X-axis) In each subcaption, we report values for α. Coloured lines are the averages. The confidence intervals around them are calculated at the 0.95 level.



(a) $\alpha = 0.5$

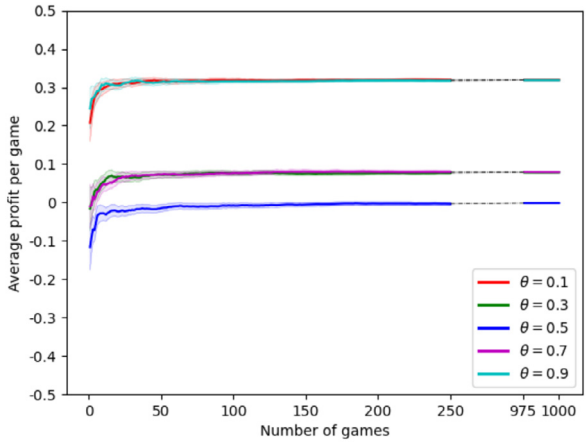


(b) $\alpha = 0.05$

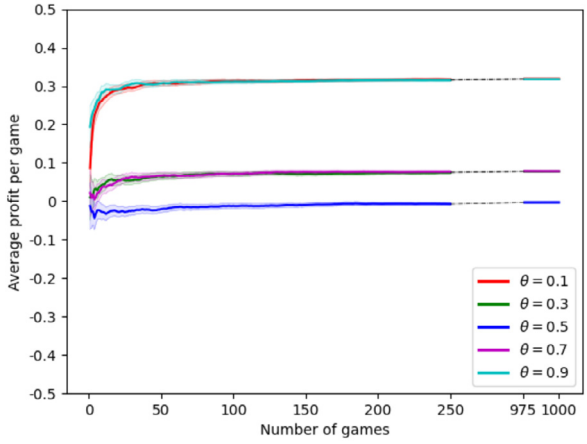


(c) $\alpha = 0.01$

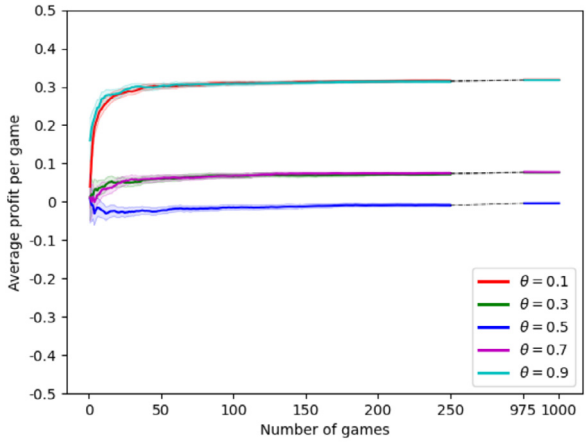
Fig. 8. Regret. Average profit per bet (Y-axis) against number of bets (X-axis) In each subcaption, we report values for α. Coloured lines are the averages. The confidence intervals around them are calculated at the 0.95 level.



(a) $\alpha = 0.5$



(b) $\alpha = 0.05$



(c) $\alpha = 0.01$

Fig. 9. ORO. Average profit per bet (Y-axis) against number of bets (X-axis) In each subcaption, we report values for α. Coloured lines are the averages. The confidence intervals around them are calculated at the 0.95 level.

References

[1] D.J. Benjamin, J.O. Berger, M. Johannesson, B.A. Nosek, E.-J. Wagenmakers, R. Berk, K.A. Bollen, B. Brembs, L. Brown, C. Camerer, et al., Redefine statistical significance, Nat. Hum. Behav. 2 (1) (2018) 6–10.
[2] G. Betz, In defence of the value free ideal, Eur. J. Philos. Sci. 3 (2) (2013) 207–220.
[3] H. Blau, Ploxoma: testbed for uncertain inference, in: D. Fisher, H.J. Lenz (Eds.), Learning from Data, in: Lecture Notes in Statistics, vol. 112, Springer, New York, 1996, pp. 47–57, Chapter 5.
[4] G. Bonanno, Game Theory, CreateSpace Independent Publishing Platform, North Charleston, SC, USA, 2015.
[5] R. Bradley, Decision Theory with a Human Face, Cambridge University Press, Cambridge, UK, 2017.
[6] S. Bradley, How to choose among choice functions, in: T. Augustin, S. Doria, E. Miranda, E. Quaeghebeur (Eds.), ISIPTA '15: Proceedings of the Ninth International Symposium on Imprecise Probability: Theories and Applications, 2015, pp. 57–66.
[7] S. Bradley, Imprecise probabilities, spring 2019 edition, in: E.N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy, Metaphysics Research Lab, Stanford University, 2019, accessed: 2023-05-01.
[8] S. Bradley, Learning by ignoring the most wrong, KRITERION–J. Philos. 36 (1) (2022) 9–31.
[9] S. Bradley, K. Steele, Can free evidence be bad? Value of information for the imprecise probabilist, Philos. Sci. 83 (1) (2016) 1–28.
[10] J. de Ruiter, Redefine or justify? Comments on the alpha debate, Psychon. Bull. Rev. 26 (2) (2018) 430–433.
[11] D. Ellsberg, Risk, ambiguity, and the Savage axioms, Q. J. Econ. 75 (4) (1961) 643–669.
[12] D. Fanelli, "Positive" results increase down the hierarchy of the sciences, PLoS ONE 5 (4) (2010) e10068.
[13] F. Fidler, J. Wilcox, Reproducibility of scientific results, in: E.N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy, summer 2021 edition, Metaphysics Research Lab, Stanford University, 2021.
[14] P.C. Fishburn, Utility Theory for Decision Making, John Wiley and Sons, New York, 1970.
[15] R. Gong, J.B. Kadane, M.J. Schervish, T. Seidenfeld, R.B. Stern, Learning and total evidence with imprecise probabilities, Int. J. Approx. Reason. 151 (2022) 21–32.
[16] A. Greenwald, R. Gonzalez, R. Harris, D. Guthrie, Effect sizes and p values: what should be reported and what should be replicated?, Psychophysiology 33 (2) (1996) 175–183.
[17] R. Haenni, J.-W. Romeijn, G. Wheeler, J. Williamson, Probabilistic Logics and Probabilistic Networks, Springer, Dordrecht, the Netherlands, 2011.
[18] J. Hawthorne, J. Landes, C. Wallmann, J. Williamson, The principal principle implies the principle of indifference, Br. J. Philos. Sci. 68 (1) (2017) 123–131.
[19] J.L. Hodges Jr., E.L. Lehmann, The use of previous experience in reaching statistical decisions, Ann. Math. Stat. 23 (3) (1952) 396–407.
[20] L. Hurwicz, The generalised Bayes-minimax principle: a criterion for decision-making under uncertainty, in: Cowles Commission Discussion Paper: Statistics, vol. 355, 1951, pp. 1–7.
[21] E.T. Jaynes, Probability Theory, Cambridge University Press, Cambridge, UK, 2003, Edited version by G. Larry Bretthorst.
[22] H. Jeffreys, The Theory of Probability, Oxford University Press, Oxford, UK, 1998.
[23] V.E. Johnson, Revised standards for statistical evidence, Proc. Natl. Acad. Sci. 110 (48) (2013) 19313–19317.
[24] L.V. Kantorovich, Mathematical methods of organizing and planning production, Manag. Sci. 6 (4) (1960) 366–422.
[25] J.M. Keynes, The general theory of employment, Q. J. Econ. 51 (2) (1937) 209–223.
[26] D. Kreps, Notes on the Theory of Choice, Routledge, New York, 1988.
[27] H.E. Kyburg, Probability and the Logic of Rational Belief, Wesleyan University Press, Middletown CT, USA, 1961.
[28] H.E. Kyburg, The Logical Foundations of Statistical Inference, Springer, Dordrecht, the Netherlands, 1974.
[29] H.E. Kyburg, Science and Reason, Oxford University Press, Oxford, UK, 1990.
[30] H.E. Kyburg, The scope of Bayesian reasoning, in: D.L. Hull, M. Forbes, K. Okruhlik (Eds.), PSA 1992: Proceedings of the Biennial Meeting of the Philosophy of Science Association, vol. 2, Cambridge University Press, 1992, pp. 139–152.
[31] H.E. Kyburg, Combinatorial semantics: semantics for frequent validity, Comput. Intell. 13 (2) (1997) 215–257.
[32] H.E. Kyburg, Are there degrees of belief?, J. Appl. Log. 1 (3–4) (2003) 139–149.
[33] H.E. Kyburg, Belief, evidence, and conditioning, Philos. Sci. 73 (1) (2006) 42–65.
[34] H.E. Kyburg, M. Pittarelli, Set-based bayesianism, IEEE Trans. Syst. Man Cybern., Part A, Syst. Hum. 26 (3) (1996) 324–339.
[35] H.E. Kyburg, C.M. Teng, Choosing among interpretations of probability, in: K.B. Laskey, H. Prade (Eds.), Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence, UAI'99, Morgan Kaufmann Publishers Inc., San Francisco CA, USA, 1999, pp. 359–365.
[36] H.E. Kyburg, C.M. Teng, Uncertain Inference, Cambridge University Press, Cambridge, UK, 2001.
[37] H.E. Kyburg, C.M. Teng, The logic of risky knowledge, reprised, Int. J. Approx. Reason. 53 (3) (2012) 274–285.
[38] H.E. Kyburg, C.M. Teng, G. Wheeler, Conditionals and consequences, J. Appl. Log. 5 (4) (2007) 638–650.
[39] J. Landes, Min–max decision rules for choice under complete uncertainty: axiomatic characterizations for preferences over utility intervals, Int. J. Approx. Reason. 55 (5) (2014) 1301–1317.
[40] J. Landes, Rules of proof for maximal entropy inference, Int. J. Approx. Reason. 153 (2023) 144–171.
[41] I. Levi, On indeterminate probabilities, J. Philos. 71 (13) (1974) 391–418.
[42] I. Levi, The Enterprise of Knowledge: An Essay on Knowledge, Credal Probability, and Chance, MIT Press, 1980.
[43] I. Levi, The paradoxes of Allais and Ellsberg, Econ. Philos. 2 (1) (1986) 23–53.
[44] I. Levi, Inductivism and Parmenidean epistemology: Kyburg's way, Int. J. Approx. Reason. 53 (3) (2012) 286–292.
[45] G. Masterton, Equivocation for the objective Bayesian, Erkenntnis 80 (2) (2015) 403–432.
[46] P. McCorduck, C. Cfe, Machines Who Think: A Personal Inquiry into the History and Prospects of Artificial Intelligence, CRC Press, New York NY, USA, 2004.
[47] J. Paris, A. Vencovská, G. Wilmers, A natural prior probability distribution derived from the propositional calculus, Ann. Pure Appl. Log. 70 (3) (1994) 243–285.
[48] W. Peden, Evidentialism, inertia, and imprecise probability, Br. J. Philos. Sci. (2022) 1–23.
[49] A. Perea, Epistemic Game Theory: Reasoning and Choice, Cambridge University Press, New York, NY, USA, 2012.
[50] M. Radzvilas, W. Peden, F. De Pretis, A battle in the statistics wars: a simulation-based comparison of Bayesian, Frequentist and Williamsonian methodologies, Synthese 199 (5–6) (2021) 13689–13748.
[51] M.D. Resnik, Choices: An Introduction to Decision Theory, University of Minnesota Press, Minneapolis MN, USA, 1987.
[52] L.J. Savage, The theory of statistical decision, J. Am. Stat. Assoc. 46 (253) (1951) 55–67.
[53] S. Seabold, J. Perktold, Statsmodels: econometric and statistical modeling with Python, in: Stéfan van der Walt, Jarrod Millman (Eds.), Proceedings of the 9th Python in Science Conference, 2010, pp. 92–96.
[54] T. Seidenfeld, Direct inference and inverse inference, J. Philos. 75 (12) (1978) 709–730.
[55] T. Seidenfeld, A contrast between two decision rules for use with (convex) sets of probabilities: Γ-maximin versus E-admissibilty, Synthese 140 (1/2) (2004) 69–88.
[56] T. Seidenfeld, Forbidden fruit: when epistemic probability may not take a bite of the Bayesian apple, in: W. Harper, G. Wheeler (Eds.), Probability and Inference, Essays in Honour of Henry E. Kyburg Jr., in: Texts in Philosophy, vol. 2, College Publications, London, UK, 2007, pp. 267–279.

[57] J. Stoye, Minimax regret treatment choice with finite samples, J. Econom. 151 (1) (2009) 70–81.
[58] C.M. Teng, Precisely imprecise: a collection of papers dedicated to Henry E. Kyburg, Jr., Int. J. Approx. Reason. 53 (3) (2012) 273.
[59] L.N. Vaserstein, Markov processes over denumerable products of spaces, describing large systems of automata, Probl. Pereda. Inf. 5 (3) (1969) 64–72.
[60] P. Walley, Statistical Reasoning with Imprecise Probabilities, Chapman and Hall, 1991.
[61] P. Walley, Inferences from multinomial data: learning about a bag of marbles, J. R. Stat. Soc., Ser. B, Methodol. 58 (1) (1996) 3–34.
[62] G. Wheeler, J. Williamson, Evidential probability and objective Bayesian epistemology, in: P.S. Bandyopadhyay, M.R. Forster (Eds.), Philosophy of Statistics, in: Handbook of the Philosophy of Science, vol. 7, Elsevier, 2011, pp. 307–331.
[63] J. Williamson, Motivating objective Bayesianism: from empirical constraints to objective probabilities, in: W. Harper, G. Wheeler (Eds.), Probability and Inference: Essays in Honour of Henry E. Kyburg, Jr., in: Texts in Philosophy, vol. 2, College Publications, London, UK, 2007, pp. 155–183.
[64] J. Williamson, Objective Bayesian probabilistic logic, J. Algorithms 63 (4) (2008) 167–183.
[65] J. Williamson, in: Defence of Objective Bayesianism, Oxford University Press, Oxford, UK, 2010.
[66] J. Williamson, Objective Bayesianism, Bayesian conditionalisation and voluntarism, Synthese 178 (1) (2011) 67–85.
[67] J. Williamson, From Bayesian epistemology to inductive logic, J. Appl. Log. 11 (4) (2013) 468–486.
[68] J. Williamson, How uncertain do we need to be?, Erkenntnis 79 (6) (2013) 1249–1271.
[69] J. Williamson, Why frequentists and Bayesians need each other, Erkenntnis 78 (2) (2013) 293–318.