A Theory of Bounded Inductive Rationality

Caspar Oesterheld
Computer Science Department
Carnegie Mellon University
Pittsburgh, PA, USA
oesterheld@cmu.edu

Abram Demski
Machine Intelligence Research Institute
Berkeley, California, USA

Vincent Conitzer
Computer Science Department
Carnegie Mellon University
Pittsburgh, PA, USA

The dominant theories of rational choice assume logical omniscience. That is, they assume that when facing a decision problem, an agent can perform all relevant computations and determine the truth value of all relevant logical/mathematical claims. This assumption is unrealistic when, for example, we offer bets on remote digits of π or when an agent faces a computationally intractable planning problem. Furthermore, the assumption of logical omniscience creates contradictions in cases where the environment can contain descriptions of the agent itself. Importantly, strategic interactions as studied in game theory are decision problems in which a rational agent is predicted by its environment (the other players). In this paper, we develop a theory of rational decision making that does not assume logical omniscience. We consider agents who repeatedly face decision problems (including ones like betting on digits of π or games against other agents). The main contribution of this paper is to provide a sensible theory of rationality for such agents. Roughly, we require that a boundedly rational inductive agent tests each efficiently computable hypothesis infinitely often and follows those hypotheses that keep their promises of high rewards. We then prove that agents that are rational in this sense have other desirable properties. For example, they learn to value random and pseudo-random lotteries at their expected reward. Finally, we consider strategic interactions between different agents and prove a folk theorem for what strategies bounded rational inductive agents can converge to.

1 Introduction

The dominant theories of rational decision making – in particular Bayesian theories – assume logical omniscience, i.e., that rational agents can determine the truth value of any relevant logical statement. In some types of decision problems, this prevents one from deriving any recommendation from these theories, which is unsatisfactory (Sect. 3). For one, there are problems in which computing an optimal choice is simply computationally intractable. For example, many planning problems are intractable. Second, the assumption of logical omniscience creates contradictions (resembling classic paradoxes of self reference, such as the liar's paradox) if the environment is allowed to contain references to the agent itself. These issues arise most naturally when multiple rational agents interact and reason about one another.

This paper develops a novel theory of boundedly rational inductive agents (BRIAs) that does not assume logical omniscience and yields sensible recommendations in problems such as the ones described above. Rather than describing how an agent should deal with an individual decision, the theory considers how an agent learns to choose on a sequence of different decision problems. We describe the setting in more detail in Sect. 2.

The core of our theory is a normative rationality criterion for such learning agents. Roughly, the criterion requires that a boundedly rational inductive agent test each efficiently computable hypothesis (or more generally each hypothesis in some class) infinitely often and follows hypotheses that keep their promises of high rewards. We describe the criterion in detail in Sect. 4. Importantly, the criterion can be satisfied by computationally bounded agents, as we show in Sect. 5.

R. Verbrugge (Ed.): Theoretical Aspects of
Rationality and Knowledge 2023 (TARK 2023)
EPTCS 379, 2023, pp. 421–440, doi:10.4204/EPTCS.379.33

© C. Oesterheld, A. Demski & V. Conitzer
This work is licensed under the
Creative Commons Attribution License.

We demonstrate the appeal of our criterion by showing that it implies desirable and general behavioral patterns. In Sect. 6, we show that on sequences of decision problems in which one available option guarantees a payoff of at least l, BRIAs learn to obtain a reward of at least l. Thus, in particular, they avoid Dutch books (in the limit). We further show that similarly on sequences of decision problems in which one available option pays off truly or algorithmically randomly with mean μ, BRIAs learn to obtain a reward of at least μ. Finally, we consider decision problems in which one BRIA plays a strategic game against another BRIA. We show that BRIAs can converge to any individually rational correlated strategy profile. BRIAs are thus a promising model for studying ideas such as superrationality (i.e., cooperation in the one-shot Prisoner's Dilemma) [11] (cf. Sect. 8). Related work is discussed in Sect. 8. Throughout this paper, we describe the key ideas for our proofs in the main text. Detailed proofs are given in Appendix A.

2 Setting

Informally, we consider an agent who makes decisions in discrete time steps. At each time step she faces some set of available options to choose from. She selects one of options and receives a reward. She then faces a new decision problem, and so on.

Formally, let $\mathscr{T}$ be some language describing available *options*. A *decision problem* $\mathrm{DP} \in \mathrm{Fin}(\mathscr{T})$ is a finite set of options. A *decision problem sequence* is a sequence of decision problems $\mathrm{DP}_1, \mathrm{DP}_2, ...$ An agent for $\overline{\mathrm{DP}}$ is a sequence $\bar{c}$ of $c_t \in \mathrm{DP}_t$. The rewards are numbers $r_1, r_2, r_3, ... \in [0,1]$. Note that in contrast to the literature on multi-armed bandit problems (Sect. 8) counterfactual rewards are not defined.

It is generally helpful to imagine that (similar to multi-armed bandit problems) at each time t the agent first sees DP_t; then chooses c_t from DP_t according to some algorithm that looks at the available options in DP_t and takes past experiences into account; then the environment calculates some reward as a function of c_t; the agent observes the reward and learns from it. The sequence of decision problems DP_t may in turn be calculated depending on the agent's choices.

We focus on learning myopically optimal behavior. That is, we want our agent to learn to choose whatever gives the highest reward for the present decision problem, regardless of what consequences that has for future decision problems.

3 Computational constraints and paradoxes of self-reference

In this paper, we develop a normative theory of rational learning in this setting. The standard theory for rational decision making under uncertainty is Bayesian decision theory (BDT) ([24, 12]; for contemporary overviews, see [19, 29]). The main ideas of this paper are motivated by a specific shortcoming of BDT: the assumption that the agent who is subject to BDT's recommendations is logically omniscient and in particular not limited by any computational constraints. We develop a theory that gives recommendations to computationally bounded agents. In the following, we give two kinds of examples to illustrate the role of logical omniscience in BDT and motivate our search for an alternative theory.

Mere intractability The first problem is that in most realistic choice problems, it is intractable to follow BDT. Bayesian updating and Bayes-optimal decision making are only feasible if the environment is small or highly structured. Even if the agent had a perfectly accurate world model, determining the optimal choice may require solving computationally hard problems, such as the traveling salesman problem, planning in 2-player competitive games, etc. Optimal choice may also rely on whether particular mathematical claims are true, e.g., when assessing the safety of particular cryptographic methods. In all

these problems, BDT requires the agent to perfectly solve the problem at hand. However, we would like a theory of rational choice that makes recommendations to realistic, bounded agents who can only solve such problems approximately.

Consider a decision problem $\mathrm{DP} = \{a_1, a_2\}$, where the agent knows that option a_1 pays off the value of the 10^{100}-th digit of the binary representation of π. Option a_2 pays off 0.6 with certainty. In our formalism, r equals the 10^{100}-th digit of the binary representation of π if $c = a_1$ and $r = 0.6$ if $c = 0.6$. All that Bayesian decision theory has to say about this problem is that one should calculate the 10^{100}-th digit of π; if it is 1, choose a_1; otherwise choose a_2. Unfortunately, calculating the 10^{100}-th digit of π is likely intractable.[1] Hence, Bayesian decision theory does not have any recommendations for this problem for realistic reasoners. At the same time, we have the strong normative intuition that – if digits of π indeed cannot be predicted better than random under computational limitations – it is rational to take a_2. We would like our theory to make sense of that intuition.

We close with a note on what we can expect from a theory about rational decision making under computational bounds. A naïve hope might be that such a theory could tell us how to optimally use some amount of compute (say, 10 hours on a particular computer system) to approximately solve any given problem (cf. our discussion in Sect. 8 of Russell et al.'s [21, 23, 22] work on bounded optimality); or that it might tell us *in practice* at what odds to bet on, say, Goldbach's conjecture with our colleagues. In this paper, we do not provide such a theory and such a theory cannot exist. We must settle for a more modest goal. Since our agents face decision problems repeatedly, our rationality requirement will be that the agent *learns* to approximately solve these problems optimally in the limit. For example, if digits of π are pseudo-random in the relevant sense, then a rational agent must converge to betting 50-50 on remote binary digits of π. But it need not bet 50-50 "out-of-the-box".

Paradoxes of self-reference, strategic interactions, and counterfactuals A second problem with BDT and logical omniscience more generally is that it creates inconsistencies if the values of different available options depend on what the agent chooses. As an example, consider the following decision problem, which we will call the Simplified Adversarial Offer (SAO) (after a decision problem introduced by [18]). Imagine that an artificial agent chooses between two available alternatives a_0 and a_1, where a_0 is known to pay off $1/2$ with certainty, and a_1 is known to pay off 1 if the agent's program run on this decision problem chooses a_0, and 0 otherwise. Now assume that the agent chooses deterministically and optimally given a logically omniscient belief system. Then the agent knows the value of each of the options. This also means that it knows whether it will select a_0 or a_1. But given this knowledge, the agent selects a different option than what the belief system predicts. This is a contradiction. Hence, there exists no agent that complies with standard BDT in this problem. Compare the example of Oesterheld and Conitzer [18] and Spencer [28]; also see Demski and Garrabrant ([7], Sect. 2.1) for a discussion of another, subtler issue that arises from logical omniscience and introspection.

We are particularly interested in problems in which such failure modes apply. SAO is an extreme and unrealistic example, selected to be simple and illustrative. However, strategic interactions between different rational agents share the ingredients of this problem: Agent 1 is thinking about what agent 2 is choosing, thereby creating a kind of reference to agent 2 in agent 2's environment. We might even imagine that two AI players know each others' exact source code (cf. [20], Sect. 10.4; [31]; [32]; [3]; [6]; [17]). Further, it may be in agent 2's interest to prove wrong whatever agent 1 believes about agent 2. For a closely related discussion of issues of bounded rationality and the foundations of game theory,

[1] Remote digits of π are a canonical example in the literature on bounded rationality and logical uncertainty (see [25], for an early usage). To the knowledge of the authors it is unknown whether the n-th digit of π can be guessed better than random in less than $O(n)$ time. For a general, statistical discussion of the randomness of digits of π, see Marsaglia [14].

see Binmore [4] and references therein (cf. [20], Ch. 10; [7]).

4 The rationality criterion

4.1 Preliminary definitions

An *estimating agent* $\bar{\alpha}$ is a sequence of choices from the available options $\alpha_t^c \in \mathrm{DP}_t$ and *estimates* $\alpha_t^e \in [0,1]$. Our rationality criterion uses estimating agents. For brevity, we will say *agent* instead of estimating agent throughout the rest of this paper. For example, let $\mathrm{SAO}_{\alpha,t}$ be the Simplified Adversarial Offer for the agent at time t as described in Sect. 3. Then we might like an agent who learns to choose $\alpha_t^c = a_0$ (which pays $1/2$ with certainty) and estimate $\alpha_t^e = 1/2$.

A *hypothesis* h has the same type signature as an estimating agent. When talking about hypotheses, we will often refer to the values of h_t^e as promises and to the values of h_t^c as recommendations.

Our rationality criterion will be relative to a particular set of hypotheses $\mathbb{H}$. In principle, $\mathbb{H}$ could be any set of hypotheses, e.g., all computable ones, all three-layer neural nets, all 8MB computer programs, etc. Generally, $\mathbb{H}$ should contain any hypothesis (i.e., any hypothesis about how the agent should act) that the agent is willing to consider, similar to the support of the prior in Bayesian theories of learning, or the set of experts in the literature on multi-armed bandits with expert advice. Following Garrabrant et al. [10], we will often let $\mathbb{H}$ be the set of functions computable in $O(g(t))$ time, where g is a non-decreasing function. We will call these hypotheses *efficiently computable (e.c.)*. Note that not all time complexity classes can be written as $O(g(t))$. For example, the set of functions computable in polynomial time cannot be written in such a way. This simplified set is used to keep notation simple. Our results generalize to more general computational complexity classes.

4.2 No overestimation

We now describe the first part of our rationality requirement, which is that the estimates should not be systematically above what the agent actually obtains. The criterion itself is straightforward, but its significance will only become clear in the context of the hypothesis coverage criterion of the next section.

Definition 1. *For $T \in \mathbb{N}$, we call $\mathscr{L}_T(\bar{\alpha},\bar{r}) := \sum_{t=1}^{T} \alpha_t^e - r_t$ the cumulative overestimation of an agent $\bar{\alpha}$ on $\bar{r}$.*

Definition 2. *We say that an agent $\bar{\alpha}$ for $\overline{\mathrm{DP}},\bar{r}$ does not overestimate (on average in the limit) if $\mathscr{L}_T(\bar{\alpha},\bar{r})/T \leq 0$ as $T \to \infty$.*

In other words, for all $\varepsilon > 0$, there should be a time t such that for all $T > t$, $\mathscr{L}_T(\bar{\alpha},\bar{r})/T \leq \varepsilon$. Note that the per-round overestimation of boundedly rational inductive agents as defined below will usually but need not always converge to 0; it can be negative in the limit.

4.3 Covering hypotheses

We come to our second requirement, which specifies how the agent $\bar{\alpha}$ relates to the hypotheses in $\mathbb{H}$.

Definition 3. *We say that $\bar{h}$ outpromises $\bar{\alpha}$ or that $\bar{\alpha}$ rejects $\bar{h}$ at time t if $h_t^e > \alpha_t^e$.*

We distinguish two kinds of hypotheses: First, there are hypotheses that promise higher rewards than $\bar{\alpha}^e$ in only finitely many rounds. For example, this will be the case for hypotheses that $\bar{\alpha}$ trusts and takes into account when choosing and estimating. Also, this could include hypotheses who recommend an inferior option with an accurate estimate, e.g., hypotheses that recommend "$1/3$" and promise $1/3$ in

$\{$"1/3", "2/3"$\}$. For all of these hypotheses, we do not require anything of $\bar{\alpha}$. In particular, $\bar{\alpha}$ need not test these hypotheses. Second, some hypotheses do infinitely often outpromise $\bar{\alpha}^e$. For these cases, we will require our boundedly rational inductive agents to have some reason to reject these hypotheses. To be able to provide such a reason, $\bar{\alpha}$ needs to test these hypotheses infinitely often. Testing a hypothesis requires choosing the hypothesis' recommended action.

Definition 4. *We call a set $M \subseteq \mathbb{N}$ a test set of $\bar{\alpha}$ for $\bar{h}$ if for all $t \in M$, $\alpha_t^c = h_t^c$.*

For $\bar{\alpha}$ to infinitely often reject $\bar{h}$, these tests must then show that $\bar{h}$ is not to be trusted (in those rounds in which they promise a reward that exceeds $\bar{\alpha}^e$). That is, on these tests, the rewards must be significantly lower than what the hypothesis promises. We thus introduce another key concept.

Definition 5. *Let $\bar{h}$ be a hypothesis and $M \subseteq \mathbb{N}$ be a test set of $\bar{\alpha}$ for $\bar{h}$. We call $l_T(\bar{\alpha}, \bar{r}, M, \bar{h}) := \sum_{t \in M_{\leq T}} r_t - h_t^e$ the (empirical) record of h (on M).*

Here, $M_{\leq T} := \{t \in M \mid t \leq T\}$ is defined to be the set of elements of M that are at most T.

We now have all the pieces together to state the coverage criterion, which specifies how we want our agents to relate to the hypotheses under consideration.

Definition 6. *Let $\bar{\alpha}$ be an agent, $\bar{h}$ be a hypothesis, and let B be the set of times t at which $\bar{\alpha}$ rejects $\bar{h}$. We say that $\bar{\alpha}$ covers $\bar{h}$ with test set M if either B is finite or the sequence $\left(l_T(\bar{\alpha}, \bar{r}, M, \bar{h})\right)_{T \in B}$ goes to negative infinity.*

4.4 The boundedly rational inductive agent criterion

We now state the BRIA criterion, the main contribution of this paper.

Definition 7. *Let $\bar{\alpha}$ be an agent for $\overline{\mathrm{DP}}, \bar{r}$. Let $\mathbb{H} = \{h_1, h_2, \ldots\}$ be a set of hypotheses. We say $\bar{\alpha}$ is a boundedly rational inductive agent (BRIA) for $\overline{\mathrm{DP}}, \bar{r}$ covering $\mathbb{H}$ with test sets $M_1, M_2, \ldots$ if $\bar{\alpha}$ does not overestimate and for all i, $\bar{\alpha}$ covers h_i with test set M_i.*

In the following, whenever $\bar{\alpha}$ is a BRIA, we will imagine that the test sets are given as a part of $\bar{\alpha}$. For example, if we say that $\bar{\alpha}$ is computable in, say, time polynomial in t, then we will take this to mean that $\bar{\alpha}$ together with a list at time t of tested hypotheses can be computed in polynomial time.

4.5 Examples

Betting on digits of π Consider the decision problem sequence with $\mathrm{DP}_t = \{a_t^\pi, x_t\}$ for all t, where a_t^π pays off the 2^t-th binary digit of π – i.e., r_t is the 2^t-th digit of π if $\alpha_t^c = a_t^\pi$ – and $x_t \in [0,1]$ pays off x_t. As usual we assume that the 2^t-th binary digits of π are pseudorandom (in a way we will make precise in Sect. 6) uniformly distributed (as they seem to be, cf. footnote 1). We would then expect boundedly rational agents to (learn to) choose a_t^π when $x_t < 1/2$ and choose x_t when $x_t > 1/2$.

We now consider an agent $\bar{\alpha}$ for this decision problem sequence. We will step-by-step impose the components of the BRIA criterion on $\bar{\alpha}$ to demonstrate their meaning and (joint) function in this example. We start by imposing the no overestimation criterion on $\bar{\alpha}$ without any assumptions about hypothesis coverage – what can we say about $\bar{\alpha}$ if we assume that does not overestimate? As noted earlier, the no overestimation criterion alone is weak and in particular does not constrain choice at all. For instance, $\bar{\alpha}$ might always choose $\alpha_t^c = a_t^\pi$ and alternate estimates of 0 and 1; or it might always choose x_t and estimate x_{t-1}.

We now impose instances of the hypothesis coverage criterion. We start with the hypothesis h_x which always recommends choosing x_t and promises a reward of x_t. Note that for all we know about the

decision problem sequence this hypothesis does not give particularly good recommendations. However, in the context of our theory, h_x is useful because it always holds its promises. In particular, h_x's empirical record on any test set is 0. Hence, if α is to cover h_x, then α can only reject h_x finitely many times. By definition, this means that $\alpha_t^e \geq x_t$ for all but finitely many $t \in \mathbb{N}$. With the no overestimation criterion, it follows that α on average obtains utilities at least equal to x_t. But α's choices may still not match our bounded ideal. For example, α may always choose x_t.

Next, consider for $\varepsilon > 0$, the hypothesis h_π^ε that always recommends a_t^π and estimates $1/2 - \varepsilon$. Whether h_π^ε holds its promises is a more complicated question. But let us assume that $\bar{\alpha}$ covers h_π^ε with some test set M, and let us further assume that whether $t \in M$ is uncorrelated with the 2^t-th binary digit of π, for instance, because predicting the 2^t-th binary digit of π better than random cannot be done using the agent's computational capabilities. Then h_π^ε's empirical record on M will go to ∞, assuming that M is infinite – after all, following h_π^ε's recommendations yields a reward of $1/2$ on average, exceeding its promises of $1/2 - \varepsilon$. With the assumption that $\bar{\alpha}$ covers h_π^ε, it follows that for all but finitely many t, $\alpha_t^e \geq 1/2 - \varepsilon$. Now imagine that α not only covers one particular h_π^ε, but that there exist arbitrarily small positive ε such that α covers the hypothesis h_π^ε. Then it follows that in the limit as $t \to \infty$, $\alpha_t^e \geq 1/2$.

The above three conditions – no overestimation, coverage of h_x and coverage of h_π^ε for arbitrarily small ε – jointly imply that $\bar{\alpha}$ exhibits the desired behavior. Specifically, we have shown that $\bar{\alpha}$ must estimate at least $\max\{1/2, x_t\}$ in the limit. By the no overestimation criterion, $\bar{\alpha}$ also has to actually obtain at least $\max\{1/2, x_t\}$ on average. And if $\bar{\alpha}$ cannot guess the 2^t-th digits of π better than random, then the only way to achieve $\max\{1/2, x_t\}$ on average is to follow with limit frequency 1 the policy of choosing a_t^π when $x_t < 1/2$ and x_t when $x_t > 1/2$.

Adversarial offers Let α be an agent who faces a sequence of instances of SAO. In particular at time t, the agent faces $\text{SAO}_{\alpha,t} = \{a_0, a_1\}$, where a_0 pays off $1/2$ with certainty. Intuitively, a_1 is evaluated to 1 if on the present problem α chooses a_0 and to 0 otherwise. Note, however, that the former fact is never relevant to computing r_t. So effectively $r_t = 1/2$ if $\alpha_t^c = a_0$ and $r_t = 0$ otherwise.

Assume that α does not overestimate and that it covers the hypothesis h which estimates $1/2$ and recommends a_0 in every round. Hypothesis h will always have an empirical record of 0 on any test set M since it holds its promises exactly. Hence, if α is to cover h, it can reject h only finitely many times. Thus, $\alpha_t^e \geq 1/2$ in all but finitely many rounds. To satisfy the no overestimation criterion, α must therefore obtain rewards of at least $1/2$ on average in the limit. Since a_1 pays off 0 whenever it is taken by α, it must be $\alpha_t^c = a_0$ with limit frequency 1.

5 Computing boundedly rational inductive agents

As described in Sect. 3, the goal of this paper is to formulate a rationality requirement that is not self-contradictory and that can be satisfied by computationally bounded agents. Therefore, we must show that one can actually construct BRIAs for given $\mathbb{H}$ and that under some assumptions about $\mathbb{H}$, such BRIAs are computable (within some asymptotic bounds).

Theorem 1. *Let $\mathbb{H}$ be a computably enumerable set consisting of ($O(g(t))$-)computable hypotheses. (Let $g \in \Omega(\log)$.) Then there exists an algorithm that computes a BRIA covering $\mathbb{H}$ (in $O(g(t)q(t))$, for arbitrarily slow-growing, $O(g(t))$-computable q with $q(t) \to \infty$) for any $\overline{\text{DP}}, \bar{r}$.*

We here give a sketch of our construction. For each decision problem, we run a first-price sealed-bid auction among the hypotheses. The highest-bidding hypothesis determines the agent's choice and estimate and is tested in this round. For each hypothesis, we maintain a wealth variable that tracks the hypothesis' empirical record. A hypothesis' bid is bound by its wealth. Thus, when a hypothesis

outpromises the agent, this implies that the hypothesis' wealth is low. Upon winning an auction, the hypothesis pays its promise and gains the reward obtained after following the hypothesis' recommendation. We further distribute at each time t allowance to the hypotheses. The overall allowance per round is finite and goes to zero. The cumulative allowance for each hypothesis goes to ∞ over time. Thus, if a hypothesis is rejected infinitely often, then this requires the hypothesis to have spent all allowance and thus for its record among those rejection rounds to go to $-\infty$. Moreover, the cumulative overestimation is bound by overall allowance distributed and thus per-round overestimation goes to 0.

The next result shows that the BRIAs given by Theorem 1 are optimal in terms of complexity.

Theorem 2. *Let α be a BRIA for $\overline{\mathrm{DP}}, \bar{r}, \mathbb{H}$. Assume that there are infinitely many t such that $|\mathrm{DP}_t| \geq 2$ and $\alpha_t^e < 1$. If $\mathbb{H}$ is the set of ($O(g(t))$-)computable hypotheses, then α is not computable (in $O(g(t))$).*

6 Lower bounds on average rewards

Options with payoff guarantees Throughout this section, we will show that BRIAs satisfy many desiderata that one might have for rational decision makers. We start with a simple result which shows that if at each time t one of the options can be efficiently shown to have a value of at least L_t, then a BRIA will come to obtain at least L_t on average.

Theorem 3. *Let $\bar{\alpha}$ be a BRIA for $\overline{\mathrm{DP}}, \bar{r}$ and the set of e.c. hypotheses. Let $\bar{a}$ be a sequence of terms in $\mathscr{T}$ s.t. for all $t \in \mathbb{N}$, it holds that $a_t \in \mathrm{DP}_t$ and $\alpha_t^c = a_t \implies r_t \geq L_t$ for some e.c. sequence $\bar{L}$. We require also that the a_t are efficiently identifiable from the sets DP_t. Then in the limit as $T \to \infty$ it holds that $\sum_{t=1}^T r_t/T \geq \sum_{t=1}^T L_t/T$.*

The proof idea is simple. Consider the hypothesis that estimates L_t and recommends a_t if $t \in S$ and promises 0 otherwise. This hypothesis always keeps its promises. Hence, to cover this hypothesis, α can be outpromised by this hypothesis only finitely many times.

We can interpret Theorem 3 as providing an immunity to money extraction schemes, a widely discussed rationality condition. If a BRIA can leave with a certain payoff of L_t, it will on average leave with at least L_t. For example, in SAO of Sect. 3, a BRIA walks away with at least $1/2$, which in turn means that it chooses $a_0 =$ "$1/2$" with frequency 1.

Options with algorithmically random payoffs Theorem 8 only tells us something about *truly* random variables. But a key goal of our theory is to also be able to assign expected rewards to *algorithmically* random sequences, i.e., sequences that are deterministic, but relevantly unpredictable under computational constraints. We first offer a formal notion of algorithmic randomness.

Definition 8. *We say a sequence $\bar{y}$ is ($O(h(t))$ boundedly) van Mises–Wald–Church (vMWC) random with means $\bar{\mu}$ if for every infinite set $S \subseteq \mathbb{N}$ that is decidable (in $O(h(t))$ time) from available information, we have that $\lim_{T\to\infty} \sum_{t\in S_{\leq T}} y_t - \mu_t = 0$.*

Thus, we call a sequence random if there is no ($O(g(t))$-)computable way of selecting in advance members of the sequence whose average differs from the means $\bar{\mu}$. Definition 8 generalizes the standard definition of (unbounded) vMWC randomness (e.g. [8], Definition 7.4.1) to non-binary values with means $\bar{\mu}$ other than $1/2$ and computational constraints with outside input (e.g., from $\overline{\mathrm{DP}}$, which could contain options containing information such as, "by the way, the trillionth digit of π is 2"). The notion of vMWC randomness is generally considered quite weak (e.g. [8], Sect. 6.2).

Theorem 4. *Let $\bar{\mu}$ be an e.c. sequence on $[0,1]$. Let α be an $O(h(t))$-computable BRIA for decision problem sequence $\overline{\mathrm{DP}}$ with rewards $\bar{r}$ covering all e.c. hypotheses. Let $\bar{a}$ be a sequence of terms in $\mathscr{T}$ s.t. $a_t \in \mathrm{DP}_t$ for all $t \in \mathbb{N}$ and the payoffs r_t in rounds with $\alpha_t^c = a_t$ are $O(h(t))$-boundedly vMWC random with means $\bar{\mu}$. Then in the limit as $T \to \infty$, it holds that $\sum_{t=1}^T r_t/T \geq \sum_{t=1}^T \mu_t/T$.*

We show an analogous result for Schnorr bounded randomness [26, 1, 33, 30] in Appendix E. Analogous results for truly random options follow from results for algorithmically random $\bar{r}$ and the fact that a sequence of truly random, independent numbers is algorithmically random almost surely. We give a direct proof in Appendix B.

7 Boundedly rational inductive agents as a foundation for game theory

We first recap basic game-theoretic concepts. A *(two-player) game* consists of two finite sets of *(pure) strategies* A_1, A_2, one set for each player, and two payoff functions $u_1, u_2 : A_1 \times A_2 \to [0,1]$. A correlated strategy profile is a distribution $\mathbf{c} \in \Delta(A_1 \times A_2)$ over $A_1 \times A_2$. We can naturally extend utility functions to correlated strategy profiles as follows: $u_i(\mathbf{c}) = \sum_{\mathbf{a} \in A_1 \times A_2} c_{\mathbf{a}} u_i(\mathbf{a})$. We call a correlated strategy profile $\mathbf{c}$ *strictly individually rational* if each player's payoff in $\mathbf{c}$ is greater than their pure strategy maximin payoff, i.e., $u_i(\mathbf{c}) > \max_{a_i \in A_i} \min_{a_{-i} \in A_{-i}} u_i(a_i, a_{-i})$.

Now imagine that two BRIAs $\bar{\alpha}_1, \bar{\alpha}_2$ learn to play a game against each other. That is, we consider BRIAs $\bar{\alpha}_1, \bar{\alpha}_2$ for $\overline{\mathrm{DP}}^{\bar{\alpha}_1}, \overline{\mathrm{DP}}^{\bar{\alpha}_2}$ respectively, where $\overline{\mathrm{DP}}^{\bar{\alpha}_i} = A_i$ for $i = 1, 2$ and $r_{i,t} = u_i(\alpha^c_{1,t}, \alpha^c_{1,t})$.

Theorem 5 (Folk theorem)**.** *Let Γ be a game. Let $\mathbb{H}_1, \mathbb{H}_2$ be any sets of hypotheses. Let $\mathbf{c} \in \Delta(A_1 \times A_2)$ be strictly individually rational. Then there exists $\mathbf{c}'$ arbitrarily close to $\mathbf{c}$ and BRIAs $\bar{\alpha}_1, \bar{\alpha}_2$ covering $\mathbb{H}_1, \mathbb{H}_2$ for decision problem sequences $\overline{\mathrm{DP}}^{\alpha_1}, \overline{\mathrm{DP}}^{\alpha_2}$ with rewards $\bar{r}_1, \bar{r}_2$ based on Γ as defined above s.t. the empirical distribution of (α^c_1, α^c_2) converges to $\mathbf{c}'$, i.e., for all $\mathbf{a} \in A_1 \times A_2$, $1/T \sum_{t=1}^{T} \mathbb{1}[(\alpha^c_1, \alpha^c_2) = \mathbf{a}] \to c'_{\mathbf{a}}$ as $T \to \infty$. Conversely, if α_1, α_2 are BRIAs for sets of hypotheses $\mathbb{H}_1$ and $\mathbb{H}_2$ that contain at least the constant-time deterministic hypotheses, $\sum_{t=1}^{T} u_i(\alpha^c_{1,t}, \alpha^c_{2,t})/T \geq \max_{a_i} \min_{a_{-i}} u_i(a_i, a_{-i})$ as $T \to \infty$. That is, in the limit each player receives at least their maximin utility.*

Theorem 5 is compelling, because it means BRIAs can learn to cooperate in one-shot games where rational agents would otherwise fail to cooperate (e.g., contrast fictitious play, or regret learning, both of which necessarily converge to defecting in the Prisoner's Dilemma). Note that our BRIA criterion is myopic, i.e., aimed at maximizing reward in the *current* round. Thus, even though the BRIAs in the above setting play repeatedly, the above result is unrelated to the folk theorems for repeated games.

8 Related work

Multi-armed bandit problems Our setting resembles a multi-armed bandit problem with expert advice (where $\mathbb{H}$ is the set of "experts"). The main difference is that we only define r_t, the reward actually obtained by the agent. The literature on multi-armed bandit problems assumes that the problem also defines the (counterfactual) rewards of untaken options and defines rationality in terms of these rewards. As discussed in Sect. 3, one of our motivations is to do away with these counterfactuals.

Within the multi-armed bandit literature, the most closely related strand of work is the literature on adversarial multi-armed bandit problems with expert advice ([2]; [13]). Like this paper, this literature addresses this problem of bounded rationality by formulating rationality relative to a set of hypotheses (the eponymous experts). However, its rationality criterion is very different from ours: they require regret minimization and in particular that cumulative regret is sublinear, a condition sometimes called Hannan-consistency. As the Simplified Adversarial Offer shows, Hannan-consistency is not achievable in our setting. However, it does become achievable if we assume that the agent has access to a source of random noise that is independent from $\overline{\mathrm{DP}}$ (see, e.g, the Exp4 algorithm of [2], Sect. 7). Importantly, the rationality criterion itself ignores the ability to randomize, i.e., it does not prescribe that the use of randomization be optimal in any sense.

We find it implausible to *require* rational agents to randomize to minimize regret; most importantly, regret minimization can require minimizing the rewards one actually obtains – see Appendix C.

Decision theory of Newcomb-like problems Problems in which the environment explicitly predicts the agent have been discussed as Newcomb-like problems by (philosophical) decision theorists ([16]). Most of this literature has focused on discussing relatively simple cases (similar to SAO). In these cases, BRIAs generally side with what has been called evidential decision theory. For example, by Theorem 3, BRIAs learn to one-box in Newcomb's problem. Of course, BRIAs differ structurally from how a decision theorist would usually conceive of an evidential decision theory-based agent. E.g., BRIAs are not based on expected utility maximization (though they implement it when feasible; see Appendix B). We also note that the decision theory literature has, to our knowledge, not produced any formal account of how to assign the required conditional probabilities in Newcomb-like problems.

Bounded rationality The motivations of the present work as per Sect. 3, especially Sect. 3, coincide with some of the motivations for the study of bounded rationality. However, other motivations have been given for the study of bounded rationality as well (see, e.g., [27], Sect. 2). More importantly, since much of bounded rationality is geared towards explaining or prescribing *human* (as opposed to AI) behavior, the characterization and analysis of "computational capacities" often differ from ours (e.g. [5]). For instance, for most humans dividing 1 by 17 is a challenge, while such calculation are trivial for computers. A few authors have also explicitly connected the general motivations of bounded rationality with paradoxes of self reference and game theory as discussed in Sect. 3 ([4], [20, Ch. 10]). Anyway, the literature on bounded rationality is vast and diverse. Much of it is so different from the present work that a comparison hardly makes sense. Below we discuss a few approaches in this literature that somewhat resemble ours. In particular, like the present paper (and Hannan consistency) they specify rationality relative to a given set of hypotheses (that in turn is defined by computational constraints).

Russell et al.'s bounded optimality Like our approach and the other approaches discussed in this related work section, Russell et al. define *bounded optimality* as a criterion relative to a set of (computationally bounded) hypotheses called *agent programs* ([21], Sect. 1.4; [23]; [22]). Roughly, an agent program is boundedly optimal if it is the optimal program from some set of bounded programs. The main difference between our and Russell et al.'s approach is that we address the problems of Sect. 3 by developing a theory of learning to make such decisions, while Russell et al. address them by moving the decision problem one level up, from the agent to the design of the agent (cf. [7], Sect. 2.2 for a discussion of this move). As one consequence, we can design general BRIAs, while it is in general hard to design boundedly optimal agents. Of course, the feasibility of designing BRIAs comes at the cost of our agents only behaving reasonably in the limit. Moreover, the designer of boundedly optimal agents as per Russell et al. may become a subject of the paradoxes of Sect. 3 in problematic ways.

Garrabrant inductors The present is in part inspired by the work of Garrabrant et al. [10], who address the problem of assigning probabilities under computational constraints and possibilities of self-reference. As an alternative to the present theory of BRIAs, one could also try to develop a theory of boundedly rational choice by maximizing expected utility using the Garrabrant inductor's probability distributions. Unfortunately, this approach fails for reasons related to the challenge of making counterfactual claims, as pointed out by Garrabrant [9]. As in the case of Hannan consistency, we can address this problem using randomization over actions. However, like Garrabrant (ibid.), we do not find it satisfactory to *require* randomization (cf. again Appendix C). We conjecture that Garrabrant inductors with (pseudo-)randomization could be used to construct BRIAs.

9 Conclusion

We developed BRIA theory as a theory of bounded inductive rationality. We gave results that show the normative appeal of BRIAs. Furthermore, we demonstrated the theory's utility by using it to justify Nash equilibrium play. At the same time, the ideas presented lead to various further research questions, some of which we have noted above. We here give three more that we find particularly interesting. Can we modify the BRIA requirement so that it implies coherence properties à la Garrabrant et al. [10]? Do the frequencies with which BRIAs play the given pure strategies of a game converge to mixed Nash and correlated equilibria? Can BRIA theory be used to build better real-world systems?

Acknowledgments

We thank Emery Cooper, Daniel Demski, Sam Eisenstat, Daniel Kokotajlo, Alexander Oldenziel, Nisan Stiennon, Johannes Treutlein and attendants of OptLearnMAS 2021 for helpful discussions.

References

[1] K. Ambos-Spies, S.A. Terwijn & Z. Xizhong (1997): *Resource bounded randomness and weakly complete problems*. *TCS* 172(1–2), pp. 195–207, doi:10.1016/S0304-3975(95)00260-X.

[2] P. Auer, N. Cesa-Bianchi, Y. Freund & R.E. Schapire (2002): *The nonstochastic multiarmed bandit problem*. *SIAM journal on computing* 32(1), pp. 48–77, doi:10.1137/S0097539701398375.

[3] M. Barasz, P. Christiano, B. Fallenstein, M. Herreshoff, P. LaVictoire & E. Yudkowsky (2014): *Robust Cooperation in the Prisoner's Dilemma: Program Equilibrium via Provability Logic*. Available at https://arxiv.org/abs/1401.5577.

[4] K. Binmore (1987): *Modeling Rational Players: Part I*. *Economics & Philosophy* 3(2), pp. 179–214, doi:10.1017/S0266267100002893.

[5] J. Conlisk (1996): *Why bounded rationality?* *Journal of Economic Literature* 34(2), pp. 669–700.

[6] A. Critch (2016): *Parametric Bounded Löb's Theorem and Robust Cooperation of Bounded Agents*. Available at https://arxiv.org/abs/1602.04184.

[7] A. Demski & S. Garrabrant (2020): *Embedded Agency*. Available at arxiv.org/pdf/1902.09469.pdf.

[8] R.G. Downey & D.R. Hirschfeldt (2010): *Algorithmic randomness and complexity*. Springer, doi:10.1007/978-0-387-68441-3.

[9] S. Garrabrant (2017): *Two Major Obstacles for Logical Inductor Decision Theory*. Available at alignmentforum.org/posts/5bd75cc58225bf06703753d4/.

[10] S. Garrabrant, T. Benson-Tilsen, A. Critch, N. Soares & J. Taylor (2017): *A Formal Approach to the Problem of Logical Non-Omniscience*. In J. Lang, editor: *Proceedings of the 16th Conference on Theoretical Aspects of Rationality and Knowledge (TARK 2017)*, 251, EPTCS, pp. 221–235, doi:10.4204/EPTCS.251.16. Available at https://arxiv.org/abs/1707.08747.

[11] D. Hofstadter (1983): *Dilemmas for Superrational Thinkers*. *Scientific American* 248(6).

[12] R.C. Jeffrey (1965): *The Logic of Decision*. McGraw-Hill, New York.

[13] T. Lattimore & C. Szepesvari (2020): *Bandit Algorithms*. To be published by Cambridge University Press, doi:10.1017/9781108571401. Available at tor-lattimore.com/downloads/book/book.pdf.

[14] G. Marsaglia (2005): *On the Randomness of Pi and Other Decimal Expansions*. *InterStat*.

[15] J. v. Neumann (1928): *Zur Theorie der Gesellschaftsspiele*. *Mathematische Annalen* 100(1), pp. 295–320, doi:10.1007/BF01448847.

[16] R. Nozick (1969): *Newcomb's Problem and Two Principles of Choice*. In N.R. et al., editor: *Essays in Honor of Carl G. Hempel*, Springer, pp. 114–146, doi:10.1007/978-94-017-1466-2_7.

[17] C. Oesterheld (2019): *Robust Program Equilibrium*. *Theory and Decision* 86(1), pp. 143–159, doi:10.1007/s11238-018-9679-3.

[18] C. Oesterheld & V. Conitzer (2021): *Extracting Money from Causal Decision Theorists*. *Phil Quarterly* 71(4), doi:10.1093/pq/pqaa086.

[19] M. Peterson (2009): *An Introduction to Decision Theory*. Cambridge University Press, doi:10.1017/CBO9780511800917.

[20] A. Rubinstein (1998): *Modeling Bounded Rationality*. Zeuthen Lecture Book Series, The MIT Press, doi:10.7551/mitpress/4702.001.0001.

[21] S. Russell & E. Wefald (1991): *Do the Right Thing – Studies in Limited Rationality*. The MIT Press.

[22] S.J. Russell & D. Subramanian (1995): *Provably Bounded-Optimal Agents*. *JAIR* 2, pp. 575–609, doi:10.1613/jair.133.

[23] S.J. Russell, D. Subramanian & R. Parr (1993): *Provably bounded optimal agents*. In: *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence (IJCAI'93)*, pp. 338–344.

[24] L.J. Savage (1954): *The Foundations of Statistics*. John Wiley and Sons, New York.

[25] L.J. Savage (1967): *Difficulties in the Theory of Personal Probability*. *Philosophy of Science* 34(4), pp. 305–310, doi:10.1086/288168.

[26] C.P. Schnorr (1971): *Zufälligkeit und Wahrscheinlichkeit. Eine algorithmische Begründung der Wahrscheinlichkeitstheorie*. *Lecture Notes in Mathematics* 218, Springer, doi:10.1007/BFb0112468.

[27] R. Selten (1990): *Bounded Rationality*. *Journal of Institutional and Theoretical Economics* 146(4).

[28] J. Spencer (2021): *An argument against causal decision theory*. *Analysis* 81(1), pp. 52–61, doi:10.1093/analys/anaa037.

[29] K. Steele & H.O. Stefánsson (2016): *Decision Theory*. In E.N. Zalta, editor: *The Stanford Encyclopedia of Philosophy*, winter 2016 edition.

[30] D.M. Stull (2020): *Resource bounded randomness and its applications*. In: *Algorithmic Randomness – Progress and Prospects*, Lecture Notes in Logic, Cambridge University Press, pp. 301–348, doi:10.1017/9781108781718.010.

[31] M. Tennenholtz (2004): *Program equilibrium*. *Games and Economic Behavior* 49(2), pp. 363–373, doi:10.1016/j.geb.2004.02.002.

[32] W. van der Hoek, C. Witteveen & M. Wooldridge (2013): *Program equilibrium – a program reasoning approach*. *International Journal of Game Theory* 42, pp. 639–671, doi:10.1007/s00182-011-0314-6.

[33] Y. Wang (2000): *Resource bounded randomness and computational complexity*. *TCS* 237(1–2), pp. 33–55, doi:10.1016/S0304-3975(98)00119-4.

A Proofs

A.1 An easy lemma about test sets

We start with a simple lemma which we will use to simplify a few of our proofs. Roughly, the lemma shows that to cover a hypothesis h, it never helps to test h in rounds in which $h_t = 0$, i.e., in rounds in which h doesn't make any promises.

Lemma 6. *Let $\bar{h}$ be a hypothesis and $N \subseteq \mathbb{N}$ s.t. $t \in N$ implies $h_t^e = 0$. Then if $\bar{\alpha}$ covers $\bar{h}$ with test set M, $\bar{\alpha}$ covers $\bar{h}$ with test set $M - N$.*

Proof. For all T, we have that

$$\begin{aligned} l_T(\bar{\alpha},\bar{r},M,\bar{h}) = \sum_{t\in M_{\leq T}} r_t - h_t^e &= \sum_{t\in M_{\leq T}-N} r_t - h_t^e + \sum_{t\in M_{\leq T}\cap N} r_t - h_t^e \\ &= \sum_{t\in M_{\leq T}-N} r_t - h_t^e + \sum_{t\in M_{\leq T}\cap N} r_t \\ &\geq \sum_{t\in M_{\leq T}-N} r_t - h_t^e \\ &= l_t(\bar{\alpha},\bar{r},M-N,\bar{h}). \end{aligned}$$

Thus, if $l_T(\bar{\alpha},\bar{r},M,\bar{h}) \to -\infty$ as $T \to -\infty$, it must also be $l_T(\bar{\alpha},\bar{r},M-N,\bar{h}) \to -\infty$ as $T \to -\infty$. □

A.2 Proof of Theorem 1

Theorem 1. *Let $\mathbb{H}$ be a computably enumerable set consisting of ($O(g(t))$-)computable hypotheses. (Let $g \in \Omega(\log)$.) Then there exists an algorithm that computes a BRIA covering $\mathbb{H}$ (in $O(g(t)q(t))$, for arbitrarily slow-growing, $O(g(t))$-computable q with $q(t) \to \infty$) for any $\overline{\mathrm{DP}}, \bar{r}$.*

Proof. Our proof is divided into four parts. First, we give the generic construction for a BRIA (1). Then we show that this is indeed a BRIA by proving that it satisfies the no overestimation criterion (2), as well as the coverage criterion (3). Finally, we show that under the assumptions stated in the theorem, this BRIA is computable in the claimed time complexity (4).

1. The construction

First, we need an *allowance function* $A : \mathbb{N} \times \mathbb{N} \to \mathbb{R}_{\geq 0}$, which for each time n, specifies a positive amount $A(n,i)$ given to hypothesis h_i's wealth at time n. The allowance function must satisfy the following requirements:

- Each hypothesis must get infinite overall allowance, i.e., $\sum_{n=1}^{\infty} A(n,i) = \infty$ for all hypotheses h_i.
- The overall allowance distributed per round n must go to zero, i.e.,

$$\sum_{n=1}^{N} \frac{1}{N} \sum_{i=1}^{\infty} A(n,i) \underset{N\to\infty}{\to} 0. \tag{1}$$

  In particular, the allowance distributed in any particular round must be finite.

An example of such a function is $A(n,i) = n^{-1}i^{-2}$.

We can finally give the algorithm itself. Initialize the wealth variables as (for example) $w_0(i) \leftarrow 0$ for each hypothesis $h_i \in \mathbb{H}$.

At time t, we run a (first-price sealed-bid[2]) auction for the present decision problem among all hypotheses. That is, we determine a winning hypothesis

$$i_t^* \in \arg\max_{i\in\mathbb{N}} \min(h_{i,t}^e, w_t(i)) \tag{2}$$

with arbitrary tie breaking. Intuitively, each hypothesis h_i bids $h_{i,t}^e$, except that it is constrained by its wealth $w_t(i)$. The idea is that if h_i has performed poorly relative to its promises, then α should not trust

[2]This format is mainly chosen for its simplicity. We could just as well use a second-price (or third-price, etc.) auction. We could use even different formats to get somewhat different BRIA-like properties. For instance, with combinatorial auctions, one could achieve cross-decision optimization.

h_i's promise for the present problem. Let $e_t^* \in [0,1]$ be the maximum (wealth-bounded) bid itself. We then define our agent at time t as $\alpha_t := (h^c_{i_t^*,t}, e_t^*)$.

We update the wealth variables as follows. For all hypotheses $i \neq i_t^*$, we merely give allowance, i.e., $w_{t+1}(i) \leftarrow w_t(i) + A(t,i)$. For the winning hypothesis i_t^*, we update wealth according to $w_{t+1}(i_t^*) \leftarrow w_t(i_t^*) + A(t,i_t^*) + r_t - e_t^*$. That is, the highest-bidding hypothesis receives the allowance and the reward obtained after following its recommendation (r_t), but pays its (wealth-bounded) bid (e_t^*).

2. No overestimation We will show that the cumulative overestimation is bounded by the sum of the allowance.

For each T, let B_T^+ be the set of hypotheses whose wealth $w_t(i)$ is positive for at least one time $t \in \{0,...,T\}$. Note that all highest-bidding hypotheses in rounds $1....,T$ are in B_T^+ for all j. We can then write the overall wealth of the hypotheses in B_T^+ at time T as

$$\sum_{i \in B_T^+} w_T(i) = \sum_{i \in B_T^+} \sum_{n=1}^{T} A(n,i) + \sum_{t=1}^{T} r_t - \alpha_t^e.$$

That is, the overall wealth at time T is the allowance distributed at times $1,...,T$ plus the money earned/lost by the highest-bidding hypotheses.

Now notice that by the construction above, if a wealth variable $w_t(i)$ is non-negative once, it remains non-negative for all future t. Thus, for all $i \in B_T^+$, $w_T(i) \geq 0$. Second, the last term is the negated cumulative overestimation of $\bar{\alpha}$. Thus, re-arranging these terms and dividing by T gives us the following upper bound on the per-round overestimation:

$$\frac{1}{T}\mathscr{L}_T(\alpha,\bar{r}) = \frac{1}{T}\left(\sum_{i \in B_T^+} \sum_{n=1}^{T} A(n,i) - \sum_{i \in B_T^+} w_T(i)\right) \leq \frac{1}{T} \sum_{i \in B_T^+} \sum_{n=1}^{T} A(n,i) \leq \sum_{i=1}^{\infty} \frac{1}{T} \sum_{n=1}^{T} A(n,i),$$

which goes to zero as $T \to \infty$ by our requirement on the function A (line 1).

3. Hypothesis coverage Given a hypothesis h_i that strictly outpromises $\bar{\alpha}$ infinitely often, we use as a test M_i, the set of times t at which h_i is the winning hypothesis (i.e., the set of times t s.t. $i = i_t^*$). We have to show that M_i is infinite, is a valid test set (as per Definition 4), and that it satisfies the justified rejection requirement in the hypothesis coverage criterion.

A) We show that M_i is infinite. That is, we need to show that infinitely often h_i is the highest-bidding hypothesis in the auction that computes $\bar{\alpha}$. Assume for contradiction that M_i is finite. We will show that at some point h_i's bidding in the construction of $\bar{\alpha}$ will not be constrained anymore by h's wealth. We will then find a contradiction with the assumption that h_i strictly outpromises α infinitely often.

Consider that for $T' > T$, it is $w_{T'}(i) = w_T(i) + \sum_{t=T+1}^{T'} A(t,i)$. That is, from time T to any time T', hypothesis i's wealth only changes by h_i receiving allowance, because i is (by assumption) not the winning hypothesis i_t^* in any round $t \geq T$. Because we required $\sum_{n=1}^{\infty} A(n,i) = \infty$, we can select a time $T* \geq T$ such that $w_{T*}(i) \geq 1$. Note that again it is also for all $t > T*$ the case that $w_t(i) \geq 1$.

We now see that if $t \geq T*$ the wealth constraints is not restrictive. That is, for all such t it is $\min(h_{i,t}^e, w_t(i)) = h_{i,t}^e$. But it is infinitely often $h_{i,t}^e > \alpha_t^e$. This contradicts the fact that by construction, α_t is equal to the highest wealth-restricted hypothesis.

B) The fact that M_i is a valid test set follows immediately from the construction – α always chooses the recommendation of the highest-bidding hypothesis.

C) We come to the justification part of the coverage criterion. Let B_i be the set of rounds in which $\bar{h}_i$ strictly outpromises $\bar{\alpha}$.

At each time $t \in B_i$, by construction $w_T(i,j) < h^e_{i,t}(\mathrm{DP}_T)$. We have that $h^e_{i,t}(\mathrm{DP}_T) \leq 1$ and

$$w_T(i) = \sum_{n=1}^{T} A(n,i) + \sum_{t \in M_i : t < T} r_t - h^e_{i,t}.$$

Hence, from the fact that $w_T(i) < h^e_{i,t}(\mathrm{DP}_T)$ for all $T \in B_i$, it follows that for all $T \in B_i$, it is

$$\sum_{t \in M_i : t < T} h^e_{i,t} - r_t > \sum_{n=1}^{T} A(n,i),$$

which goes to infinity as $T \to \infty$, as required.

4. Computability and computational complexity It is left to show that if $\mathbb{H}$ can be computably enumerated and consist only of ($O(g(t))$-)computable hypotheses, then we can implement the above-described BRIA for $\mathbb{H}$, $\overline{\mathrm{DP}}$,$\bar{r}$ in an algorithm (that runs in $O(g(t)q(t))$, for arbitrarily slow-growing, $O(g(t))$-computable q with $q(t) \to \infty$).

The main challenge is that the construction as described above performs at any time t, operations for all (potentially infinitely many) hypotheses. The crucial idea is that for an appropriate choice of A, we only need to keep track of a finite set of hypotheses, when calculating $\bar{\alpha}$ in the first T time steps. Each hypothesis starts with an initial wealth of 0. Then a hypothesis i can only become relevant at the first time t at which $A(t,i) > 0$. At any time t, we call such hypotheses *active*. Before that time, we do not need to compute $\bar{h}_i$ and do not need to update its wealth. By choosing a function A s.t. (in addition to the above conditions) $A(t,\cdot)$ has finite, e.c. support at each time t, we can keep the set of active hypotheses finite at any given time. (An example of such a function is $A(n,i) = n^{-1}i^{-2}$ for $i < n$ and $A(n,i) = 0$ otherwise.) We have thus shown that it is enough to keep track at any given time of only a finite number of hypotheses.

At any time, we therefore only need to keep track of a finite number of wealth variables, only need to compute the recommendations and promises of a finite set of hypotheses, and only need to compute a minimum of a finite set in line 2.

Computability is therefore proven. We proceed to show the claim about computational complexity. At any time t, let $C_{\max}(t)$ be the largest constant by which the computational complexity of hypotheses at time t are bounded relative to $g(t)$. Further, let $h_b(t)$ be the set of active hypotheses. Then the computational cost from simulating all active hypotheses at time t is at most $h_b(t)C_{\max}(t)g(t)$. All of $C_{\max}(t)$ and $h_b(t)$ must go to ∞ as $t \to \infty$. However, this can happen arbitrarily slowly, up to the limits of fast ($O(g(t))$) computation. Hence, if we let $q(t) = h_b(t)C_{\max}(t)g(t)$, we can let q grow arbitrarily slowly (again, up to the limits of fast computation).

Finally, we have to verify that all other calculations can be done in $O(q(t)g(t))$: To determine the winning hypothesis given everyone's promises, we have to calculate the maximum of $h_b(t) \in O(q(t))$ numbers, which can be done in $O(q(t))$ time. We also need to conduct the wealth variable updates themselves, which accounts for $O(h_b(t))$ additions. Again, this is in $O(g(t)q(t))$. And so on. □

A.3 Proof of Theorem 2

Theorem 2. *Let α be a BRIA for $\overline{\mathrm{DP}}, \bar{r}, \mathbb{H}$. Assume that there are infinitely many t such that $|\mathrm{DP}_t| \geq 2$ and $\alpha^e_t < 1$. If $\mathbb{H}$ is the set of ($O(g(t))$-)computable hypotheses, then α is not computable (in $O(g(t))$).*

This is shown by a simple diagonalization argument. If a BRIA α were computable (in $O(g(t))$), then consider the hypothesis who in rounds in which $|\mathrm{DP}_t| \geq 2$ and $\alpha_t^e < 1$, promises 1 and recommends an option other than α_t^c; and promises 0 otherwise. This hypothesis strictly outpromises α infinitely often, is computable (in $O(g(t))$) but is never tested .

A.4 Proof of Theorem 3

Theorem 3. *Let $\bar{\alpha}$ be a BRIA for $\overline{\mathrm{DP}}, \bar{r}$ and the set of e.c. hypotheses. Let $\bar{a}$ be a sequence of terms in $\mathscr{T}$ s.t. for all $t \in \mathbb{N}$, it holds that $a_t \in \mathrm{DP}_t$ and $\alpha_t^c = a_t \implies r_t \geq L_t$ for some e.c. sequence $\bar{L}$. We require also that the a_t are efficiently identifiable from the sets DP_t. Then in the limit as $T \to \infty$ it holds that $\sum_{t=1}^T r_t/T \geq \sum_{t=1}^T L_t/T$.*

Proof. We will show that if the assumptions are satisfied, then for all but finitely many t, we have that $\alpha_t^e \geq L_t$. From this and the fact that $\bar{\alpha}$ doesn't overestimate, it then follows that $\sum_{t=1}^T r_t/T \geq \sum_{t=1}^T L_t/T$.

We prove this new claim by proving a contrapositive. In particular, we assume that $\alpha_t^e < L_t$ for infinitely many t and will then show that $\bar{\alpha}$ is not a BRIA (using the other assumptions of the theorem).

Consider hypothesis $\bar{h}_i$ such that $h_{i,t} = (a_t, L_t)$. Because $\bar{L}$ is e.c. and the $\bar{a}$ are efficiently identifiable, $\bar{h}$ is e.c. We now show that $\bar{h}_i$ is not covered by $\bar{\alpha}$, which shows that $\bar{\alpha}$ is not a BRIA. By assumption, $\bar{h}_i$ strictly outpromises $\bar{\alpha}$ infinitely often. It is left to show that there is no M_i as specified in the hypothesis coverage criterion, i.e. no M_i on which $\bar{h}_i$ consistently underperforms its promises.

If $t \in M_i$, then $\alpha_t^c = h_{i,t}^c = a_t$ and therefore $r_t \geq L_t$. It follows that for all T,

$$l_T(\bar{\alpha}, \bar{r}, M_i, \bar{h}_i) = \sum_{t \in M_i : t < T} \underbrace{r_t}_{\geq L_t} - \underbrace{h_{i,t}^e}_{= L_t} \geq 0.$$

Thus, $\bar{\alpha}$ violates the coverage criterion for $\bar{h}_i$. □

A.5 Proof of Theorem 4

Definition 8. *We say a sequence $\bar{y}$ is ($O(h(t))$ boundedly) van Mises–Wald–Church (vMWC) random with means $\bar{\mu}$ if for every infinite set $S \subseteq \mathbb{N}$ that is decidable (in $O(h(t))$ time) from available information, we have that $\lim_{T\to\infty} \sum_{t \in S_{\leq T}} y_t - \mu_t = 0$.*

Theorem 4. *Let $\bar{\mu}$ be an e.c. sequence on $[0,1]$. Let α be an $O(h(t))$-computable BRIA for decision problem sequence $\overline{\mathrm{DP}}$ with rewards $\bar{r}$ covering all e.c. hypotheses. Let $\bar{a}$ be a sequence of terms in $\mathscr{T}$ s.t. $a_t \in \mathrm{DP}_t$ for all $t \in \mathbb{N}$ and the payoffs r_t in rounds with $\alpha_t^c = a_t$ are $O(h(t))$-boundedly vMWC random with means $\bar{\mu}$. Then in the limit as $T \to \infty$, it holds that $\sum_{t=1}^T r_t/T \geq \sum_{t=1}^T \mu_t/T$.*

Proof. We prove the theorem by proving that for all $\varepsilon > 0$, $\alpha_t^e \geq \mu_t - \varepsilon$ for all but finitely many t. As usual, we prove this by proving the following contrapositive: assuming this is not the case, $\bar{\alpha}$ is not a BRIA. To prove this, consider hypothesis $\bar{h}_{a,\varepsilon}$ that at each time t promises $\max(\mu_t - \varepsilon, 0)$ and recommends a_t. Since $\bar{h}_{a,\varepsilon}$ infinitely often outpromises $\bar{\alpha}$, it must tested infinitely often. Let the test set be some infinite set $M \subseteq \mathbb{N}$. By Lemma 6, we can assume WLOG that for all $t \in M$, $h_{a,\varepsilon}^e = \mu_t - \varepsilon$.

Now notice that M is by assumption computable in $O(h(t))$ given the information available at time t. Now

$$\frac{1}{|M_{i,\leq T}|} l_T(\alpha, \bar{r}, M_i, \bar{h}_i) = \frac{1}{|M_{i,\leq T}|} \sum_{t \in M_{i,\leq T}} r_t - (\mu_t - \varepsilon) \underset{\text{w.p. } 1}{\to} \varepsilon \text{ as } T \to \infty,$$

where the final step is by the fact that among rounds where $\alpha_t^c = a_t$, $\bar{r}$ is vMWC random with means $\bar{\mu}$. Hence, $\bar{h}_{a,\varepsilon}$'s record $l_T(\alpha, \bar{r}, M_i, \bar{h}_i)$ must be positive in all but finitely many rounds. Thus, $\bar{\alpha}$'s infinitely many rejections of $\bar{h}_{a,\varepsilon}$ violate the coverage criterion. □

A.6 Proof of Theorem 5

Lemma 7 (Minimax Theorem [15]). *Let (A_1, A_2, u_1, u_2) be any game. Then*

$$\max_{\sigma_i \in \Delta(A_i)} \min_{a_{-i} \in A_{-i}} u_i(\sigma_i, \sigma_{-i}) = \min_{\sigma_{-i} \in \Delta(A_{-i})} \max_{a_i \in A_i} u_i(\sigma_i, \sigma_{-i}).$$

Theorem 5 (Folk theorem). *Let Γ be a game. Let $\mathbb{H}_1, \mathbb{H}_2$ be any sets of hypotheses. Let $\mathbf{c} \in \Delta(A_1 \times A_2)$ be strictly individually rational. Then there exists $\mathbf{c}'$ arbitrarily close to $\mathbf{c}$ and BRIAs $\bar{\alpha}_1, \bar{\alpha}_2$ covering $\mathbb{H}_1, \mathbb{H}_2$ for decision problem sequences $\overline{\mathrm{DP}}^{\alpha_1}, \overline{\mathrm{DP}}^{\alpha_2}$ with rewards $\bar{r}_1, \bar{r}_2$ based on Γ as defined above s.t. the empirical distribution of (α_1^c, α_2^c) converges to $\mathbf{c}'$, i.e., for all $\mathbf{a} \in A_1 \times A_2$, $1/T \sum_{t=1}^T \mathbb{1}[(\alpha_1^c, \alpha_2^c) = \mathbf{a}] \to c'_{\mathbf{a}}$ as $T \to \infty$. Conversely, if α_1, α_2 are BRIAs for sets of hypotheses $\mathbb{H}_1$ and $\mathbb{H}_2$ that contain at least the constant-time deterministic hypotheses, $\sum_{t=1}^T u_i(\alpha_{1,t}^c, \alpha_{2,t}^c)/T \geq \max_{a_i} \min_{a_{-i}} u_i(a_i, a_{-i})$ as $T \to \infty$. That is, in the limit each player receives at least their maximin utility.*

Proof. The latter part ("Conversely,...") follows directly from Theorem 3. It is left to prove the existence claim.

We construct the BRIAs as follows. First we fix positive probabilities $p_{\mathbf{c}} \in (0,1)$ and $(p_{a_i})_{a_i \in A_i}$ for $i = 1, 2$ (WLOG assume A_1 and A_2 are disjoint) s.t. $p_{\mathbf{c}} + \sum_{i=1}^2 \sum_{a_i \in A_i} p_{a_i} = 1$. Further let v_i be some number that is strictly greater than Player i's maximin value but strictly smaller than $p_c u_i(\mathbf{c})$. By the assumption that $\mathbf{c}$ is strictly individually rational, such a number exists if we make p_c large enough. Then let $\alpha_{i,t}^e = v_i$ for all t. Then in each step the BRIAs jointly randomize[3] independently from all bidders in $\mathbb{H}_1, \mathbb{H}_2$ as follows:

- With probability $p_{\mathbf{c}}$ both players play according to $\mathbf{c}$ by jointly implementing $\mathbf{c}$, e.g., by deterministically cycling through the different strategies in the appropriate numbers.Further, $\alpha_{i,t}^e = v_i$. No hypotheses are tested.
- With probability p_{a_i}, Player i plays a_i and Player $-i$ plays from $\arg\min_{a_{-i} \in A_{-i}} u_i(a_i, a_{-i})$. Player $-i$ estimates v_{-i} and does not test any hypothesis. Player i estimates v_i and tests every hypothesis that estimates more than v_i.

We now prove that $\bar{\alpha}_1, \bar{\alpha}_2$ thus constructed are BRIAs.

No overestimation:

$$\mathscr{L}_T(\bar{\alpha}_i, \bar{r}_i)/T = \sum_{t=1}^T (\alpha_{i,t}^e - r_{i,t})/T = \sum_{t=1}^T (v_i - r_{i,t})/T \leq v_i - u_i(\mathbf{c}) \text{ as } T \to \infty.$$

By construction, $v_i - u_i(\mathbf{c}) \leq 0$.

Coverage: Let $\bar{h}_i$ be a hypothesis that outbids $\bar{\alpha}_i$ infinitely often. Then in particular $\bar{h}_i$ outbids infinitely often in rounds in which $\bar{h}_i$ recommends some a_i and $\alpha_{i,t}^c = a_i$. Thus, $\bar{h}_i$ has an infinite test set M on which the hypothesis' empirical record is

$$l_T(\bar{\alpha}_i, \bar{r}_i, M, \bar{h}_i) = \sum_{t \in M_{\leq T}} r_t - h_{i,t}^e = \sum_{t \in M_{\leq T}} \min_{a_{-i}} u_i(h_{i,t}^c, a_{-i}) - h_{i,t}^e \leq \sum_{t \in M_{\leq T}} \max_{a_i} \min_{a_{-i}} u_i(a_i, a_{-i}) - v_i \to -\infty$$

as $T \to \infty$. Thus, $\bar{h}_i$ is covered. □

[3] We here use true randomization for simplicity. The same can be achieved using algorithmic randomness.

B Options with random payoffs

The following result shows, roughly, that when choosing between different lotteries whose expected utilities are efficiently computable, BRIAs converge to choosing the lottery with the highest expected utility. When other, non-lottery options are available, BRIAs converge to performing at least as well as the best lottery option.

Theorem 8. *Let $\bar{\alpha}$ be a BRIA for $\overline{\mathrm{DP}}, \bar{r}$. Let $\bar{a}$ be a sequence of terms in $\mathscr{T}$ s.t. $a_t \in \mathrm{DP}_t$ for all $t \in \mathbb{N}$ and the values of r_t if $\alpha_t^c = a_t$ are drawn independently from distributions with e.c. means $\bar{\mu}$. Let the a_t be efficiently identifiable from DP_t. Then almost surely in the limit as $T \to \infty$, it holds that $\sum_{t=1}^{T} r_r/T \geq \sum_{t=1}^{T} \mu_t/T$.*

The proof idea similar to the proof idea for Theorem 3. It works by considering hypotheses that recommend a_t and promise $\mu_t - \varepsilon$ and noting that the empirical record of such hypotheses goes to $-\infty$ with probability 0.

Proof. We need only show that with probability 1 for all $\varepsilon > 0$ it holds that for all but finitely many times t that $\alpha_t^e \geq \mu_t - \varepsilon$. From this and the no overestimation property of $\bar{\alpha}$, the conclusion of the theorem follow.

Again we prove the following contrapositive: If there is some $\varepsilon > 0$ s.t. with some positive probability $p > 0$ we infinitely often have that $\alpha_t^e < \mu_t - \varepsilon$, then $\bar{\alpha}$ is with positive probability not a BRIA.

Consider the hypothesis $\bar{h}_{a,\varepsilon}$ that at each time t promises $\max(\mu_t - \varepsilon, 0)$ and recommends a_t. Since with probability p, $\bar{h}_{a,\varepsilon}$ infinitely often outpromises $\bar{\alpha}$, it must in these cases (and therefore with probability (at least) p) be tested infinitely often. (If not, we $\bar{\alpha}$ would in these cases not be a BRIA and we would be done.) In these cases (i.e., when $\bar{h}_{a,\varepsilon}$ is tested infinitely often), let the test set be some infinite set $M \subseteq \mathbb{N}$. (Note that M may depend on $\bar{r}$ and inherit its stochasticity. This will not matter for the following, though.) For simplicity, let M be the empty set if $\bar{h}_{a,\varepsilon}$ does not outpromise α infinitely often. By Lemma 6, we can assume WLOG that for all $t \in M$, $h_{a,\varepsilon}^e = \mu_t - \varepsilon$. Now notice that

$$\frac{1}{|M_{i,\leq T}|} l_T(\alpha, \bar{r}, M_i, \bar{h}_i) = \frac{1}{|M_{i,\leq T}|} \sum_{t \in M_{i,\leq T}} r_t - h_{a,\varepsilon,t}^e = \frac{1}{|M_{i,\leq T}|} \sum_{t \in M_{i,\leq T}} r_t - (\mu_t - \varepsilon).$$

Conditioning on the (probability p) event that h infinitely often outbids and therefore that M is infinite, it must then with probability 1 be the case that $\sum_{t \in M_{i,\leq T}} r_t - (\mu_t - \varepsilon) \underset{\text{w.p. 1}}{\to} \varepsilon$ as $T \to \infty$ by the law of large numbers. We have thus shown that with positive probability (p) $\bar{h}_{a,\varepsilon}$ outpromises $\bar{\alpha}$ infinitely often while $\bar{h}_{a,\varepsilon}$'s record $l_T(\alpha, \bar{r}, M_i, \bar{h}_i)$ is positive in all but finitely many rounds. Thus, in this positive-probability event $\bar{\alpha}$'s infinitely many rejections of $\bar{h}_{a,\varepsilon}$ violates the coverage criterion. □

C More on randomization and regret

In the literature on multi-armed bandit problems, authors usually consider the goal of regret minimization. A natural rationality requirement is for per-round average regret to go to 0. This is sometimes called Hannan consistency. For any given agent c, the Simplified Adversarial Offer SAO_c of Sect. 3 is a problem on which regret is necessarily high. However, if we assume that the agent at time t can randomize in a way that is independent of how the rewards are assigned by D_t, it can actually be ensured that per-round regret (relative to any particular hypothesis) goes to 0 (see Sect. 8).

Arguably the assumption that the agent can independently randomize is almost always satisfied for artificial agents in practice. For instance, if an agent wanted to randomize independently, then for an adversary to predict the program's choices, it would not only need to know the program's source code. It would also require (exact) knowledge of the machine state (as used by pseudo-random number generators); as well as the exact content of any stochastic input such as video streams and hardware/true random number generators. Independent randomization might not be realistic for humans (to whom randomization requires some effort), but none of these theories under discussion (the present one, regret minimization, full Bayesian updating, etc.) are directly applicable to humans, anyway.

Nevertheless, we are conceptually bothered by the assumption of independent randomization. It seems desirable for a theory of choice to make as few assumptions as possible about the given decision problems. Moreover, we can imagine situations in which independent randomization is unavailable to a given agent. It seems odd for a theory of learning to be contingent on the fact that such situations are (currently) rare or practically insignificant. A detailed discussion of this philosophical concern is beyond the scope of this paper.

In the rest of this section, we discuss the goal of regret minimization under the assumption that algorithms *can* randomize independently of $\bar{D}$. The problems discussed in this section all involve references to the agent's choice.

We consider a version of Newcomb's problem (introduced by [16]; see Sect. 8 for further discussion and references). In particular, we consider for any chooser c the decision problem $\mathrm{NP}_c = \{a_1, a_2\}$ which is resolved as follows. First, we let $D(a_1) = 1/4 + P(c = a_1)/2$. So the value of a_1 is proportional to the probability that c chooses a_1. And second, we let $D(a_2) = D(a_1) + P(c = a_1)/4$.

If we let $p = P(c = a_1)$, then the expected reward of c in this decision problem is $1/4 + p/2 + (1-p)p/4$. It is easy to see that this is strictly increasing in p and therefore maximized if $c = a_1$ deterministically. The regret, on the other hand, of c is $p^2/4$, which is also strictly increasing in p on $[0,1]$ and therefore minimized if $c = a_2$ deterministically. Similarly, the competitive ratio is given by $\frac{1/4+3p/4}{1/4+p/2+(1-p)p/4}$, which is also strictly increasing in p on $[0,1]$ and therefore also minimized if $c = a_2$ deterministically. Regret and competitive ratio minimization as rationality criteria would therefore require choosing the policy that minimizes the actual reward obtained in this scenario, only to minimize the value of actions not taken.

As noted in Sect. 8, it is a controversial among decision theorists what the rational choice in Newcomb's problem is. However, from the perspective of this paper in this particular version of the problem, it seems undesirable to require reward minimization. Also, it is easy to construct other (perhaps more convincing) cases. For example, if a high reward can be obtained by taking some action with a small probability, then regret minimizers take that action with high probability in a positive-frequency fraction of the rounds. Or consider a version of Newcomb's problem in which $D(a_1)$ is defined as before, but $D(a_2) = D(a_1)$. On such problems, Hannan-consistency is trivially satisfied by any learner, even though taking a_1 with probability 1 is clearly optimal.

D Why an even simpler theory fails and estimates are necessary

A simple mechanism of learning to choose is the *law of effect* (LoE):

> Of several responses made to the same situation, those which are accompanied or closely followed by satisfaction to the animal will, other things being equal, be more firmly connected with the situation, so that, when it recurs, they will be more likely to recur; those which are accompanied or closely followed by discomfort to the animal will, other things

> being equal, have their connections with that situation weakened, so that, when it recurs, they will be less likely to occur. The greater the satisfaction or discomfort, the greater the strengthening or weakening of the bond.

This notion is implicit in many reinforcement learning algorithms. In (human) psychology it is also known as operant conditioning.

In situations like ours, where situations generally do not repeat exactly, for the law of effect to be meaningful, we have to applied on a meta level to general hypotheses or policies for making choices. So let a policy be a function that maps observations to actions. Then we could phrase this meta LoE as: if following a particular policy is accompanied with high rewards, then an agent will follow this policy more often in the future.

The BRIA criterion can be seen as abiding by this meta LoE, as the BRIA criterion requires testing different hypotheses and following the ones that have experimentally proven themselves. Its main conceptual innovation relative to the meta LoE is the bidding system, i.e., having the agent as well as hypotheses give estimates for how much utility will be achieved by making a particular choice, and using these estimates for testing and evaluation. A natural question then is: Are these conceptual additions to meta LoE necessary to obtain the kind of results we obtain? We here show why the answer is yes.

The biggest problem is quite simple to understand: if we don't restrict the testing regimen for policies, then biased testing can justify clearly suboptimal behavior. As an illustrative example, imagine that for all t, $\mathrm{DP}_t \in \mathrm{Fin}([0,1])$ where $r_t = \alpha_t^c$. That is, at each time the agent is offered to choose from some set of numbers between 0 and 1 and then obtains as a reward the chosen number. The agent tests two policies: The first simply chooses the maximum number. The second chooses, e.g., the worst option that is greater than $1/2$ if there is one, and the best option otherwise.

Of course, in this situation one would like the agent to learn at some point to follow the max policy. BRIAs indeed learn this policy (when accompanying the two tested policies with appropriate estimates) (cf. Theorem 3). But now imagine that the agent tests the max hypothesis primarily in rounds where all values are at most $1/2$ and the other hypothesis primarily in rounds in which there are options greater than $1/2$. Then the max hypothesis could empirically be associated with lower rewards than the max hypothesis, simply because it is tested in rounds in which the maximum achievable reward is lower.

To avoid this issue we would have to require that the set of decision problems on which hypothesis A is tested is in all relevant aspects the same as the set of decision problems on which hypothesis B is tested. Unfortunately, we do not know what the "relevant aspects" are. For instance, in the above problem it may be sufficient to test the max hypothesis on even time steps and the other hypothesis on odd time steps. However, there may also be problems where rewards depend on whether the problem is faced in an even or in an odd time step. More generally, it is easy to show that for each deterministic procedure of deciding which hypothesis to test, there is a decision process $\bar{\mathrm{DP}}, \bar{r}$ in which which this testing procedure introduces a relevant bias. In particular, the positive results we have proven in Theorems 3, 4 and 8 seem out of reach. We conclude that a direct deterministic implementation of meta LoE (without the use of estimates) is insufficient for constructing a criterion of rational choice.

Besides the estimates-based approach to this problem that we have developed in this paper, a different (perhaps more obvious) approach to this problem is to test *randomly*. For this, we assume that we have a randomization device available to us that is independent of $\bar{\mathrm{DP}}, \bar{r}$. If we then, for example, randomize uniformly between testing two hypotheses, testing is unbiased in the sense that for any potentially property of decision problems, as the number of tests goes to infinity, both hypotheses will be tested on the same fraction of problems with and without that property. This is essentially the idea behind randomized controlled trials. We have discussed this idea in Appendix C.

E Schnorr bounded algorithmic randomness

Definition 9. *A martingale is a function $d\colon \mathbb{B}^* \to [0,\infty)$ s.t. for all $w \in \mathbb{B}^*$ we have that $d(w) = 1/2 d(w0) + 1/2 d(w1)$. Let $w \in \mathbb{B}^\infty$ be an infinite sequence. We say that d succeeds on w if $\limsup_{n\to\infty} d(w_1 \ldots w_n) = \infty$.*

Definition 10. *We call $w \in \mathbb{B}^\omega$ ($O(g(t))$-boundedly) Schnorr random if there is no martingale d such that d succeeds on w and d can be computed (in $O(g(t))$) given everything revealed by time t.*

Theorem 9. *Let α be an ($O(h(t))$-computable) BRIA for $\overline{\mathrm{DP}}, \bar{r}$ covering all e.c. hypotheses. Let $\bar{a}$ be a sequence of terms in $\mathscr{T}$ s.t. $a_t \in \mathrm{DP}_t$ for all $t \in \mathbb{N}$ and the values r_t in the rounds t with α_t^c are ($O(h(t))$-boundedly) Schnorr random. Then in the limit as $T \to \infty$, it holds that $\sum_{t=1}^{T} r_t/T \geq 1/2$.*

Proof. We conduct a proof by proving the following contrapositive: if the conlusion of the theorem does not hold, then $(r_t)_{t:\alpha_t^c = a_t}$ is not Schnorr random. Assume that there is $\varepsilon > 0$ s.t. $\sum_{t=1}^{T} r_t/T < 1/2 - \varepsilon$ for infinitely many T. Then by the no overestimation criterion, there must also be an $\varepsilon > 0$ s.t. $\sum_{t=1}^{T} \alpha_t^e/T < 1/2 - \varepsilon$ for infinitely many T. Consider the hypothesis $h_{a,\varepsilon}$ that always estimates $1/2 - \varepsilon$ and recommends a_t. Now let M_ε be $\bar{\alpha}$'s test for $h_{a,\varepsilon}$. From the fact that $\bar{\alpha}$ rejects $h_{a,\varepsilon}$ infinitely often, it follows that there are infinitely many $T \in \mathbb{N}$ such that $\sum_{t \in M_{\leq T}} r_t - (1/2 - \varepsilon) < 0$.

From this fact, we will now define an ($O(h(t))$-computable) martingale d that succeeds on the sequence $(r_t)_{t:\alpha_t^c = a_t}$. First, define $d() = 1$. Whenever T is not in M, define $d((r_t)_{t<T:\alpha_t^c=a_t} 0) = d((r_t)_{t<T:\alpha_t^c=a_t}) = d((r_t)_{t<T:\alpha_t^c=a_t} 1)$. That is, when $T \notin M$, don't bet on r_T. If $T \in M$, then bet some small, constant fraction δ of the current money that the next bit will be 0. That is, $d((r_t)_{t<T:\alpha_t^c=a_t} 0) = (1+\delta) d((r_t)_{t<T:\alpha_t^c=a_t})$ and $d((r_t)_{t<T:\alpha_t^c=a_t} 1) = (1-\delta) d((r_t)_{t<T:\alpha_t^c=a_t})$. Clearly, d thus defined is a martingale that is computable based on $\bar{\alpha}, M$.

Now we now know that there are infinitely many T s.t. $d((r_t)_{t<T:\alpha_t^c=a_t}) \geq (1+\delta)^{T+\varepsilon T}(1-\delta)^T$. It is left to show that for small enough δ, $(1+\delta)^{T+\varepsilon T}(1-\delta)^T \to \infty$ as $T \to \infty$.

First notice that

$$(1+\delta)^{T+\varepsilon T}(1-\delta)^T = ((1+\delta)(1-\delta))^T (1+\delta)^{T\varepsilon} = (1-\delta^2)^T (1+\delta)^{T\varepsilon} = \left((1-\delta^2)(1+\delta)^{\varepsilon}\right)^T.$$

So we need only show that for small enough but positive δ, $(1-\delta^2)(1+\delta)^\varepsilon > 1$. The most mechanic way to do this is to take the derivative at $\delta = 0$ (where the left-hand side is equal to 1) and showing that it is positive. The derivative is $\frac{d}{d\delta}(1-\delta^2)(1+\delta)^\varepsilon = (1+\delta)^\varepsilon(\varepsilon - \delta(\varepsilon+2))$. Inserting $\delta = 0$ yields ε, which is positive. □