A theory of bounded inductive rationality

Caspar Oesterheld [1] **Abram Demski** [2] **Vincent Conitzer** [3]

Abstract

The dominant theories of rational decision making assume what we will call 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 Goldbach's conjecture; 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 Goldbach's conjecture 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 show that under suitable independence assumptions, boundedly rational inductive agents can converge only to playing a Nash equilibrium against each other.

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. For one, there are problems in which computing an optimal choice is simply computationally intractable (Section 3.1). 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 (Section 3.2). These issues arise most naturally when multiple rational agents interact and reason about one another.

Drawing on ideas by Garrabrant et al. (2016), 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 Section 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 those hypotheses that keep their promises of high rewards. We describe the criterion in detail in Section 4. Importantly, the criterion can be satisfied by computationally bounded agents, as we show in Section 5.

We demonstrate the appeal of our criterion by showing that it implies various desirable and general behavioral patterns. In Section 6.1, 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). In Sections 6.2 and 6.3, we show that similarly on sequences of decision problems in which one available option pays off truly or algorithmically randomly with mean μ, BRIAs learn

[1] Computer Science Department, Carnegie Mellon University [2] Machine Intelligence Research Institute [3] Department of Computer Science, Duke University. Correspondence to: Caspar Oesterheld <caspar.oesterheld@duke.edu>.

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 under suitable independence assumptions, if BRIAs converge to some strategy profile, that strategy profile must be a Nash equilibrium. Related work is discussed in Section 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

We study a very general form of what has been called a contextual multi-armed bandit problem (see Section 8 for a discussion of that literature). Let $\mathcal{T}$ be some language describing available *options*. A *decision problem* $\mathrm{DP} \in \mathrm{Fin}(\mathcal{T})$ is a finite set of options. In this paper, we will often consider specific and somewhat unusual types of decision problems as examples, in particular ones where options are terms in some mathematical logic. However, our theory applies at least as well to more traditional, partly empirical decision problems. For example, one could imagine that each option describes a particular medical treatment and that the agent has to select one of the treatments for a particular patient.

A *decision problem sequence* consists of a sequence of decision problems $\mathrm{DP}_1, \mathrm{DP}_2, ...$ along with a sequence $\bar{D}$ of functions $D_t \colon \mathrm{DP}_t \rightsquigarrow [0,1]$ that at each time t resolve the decision problem DP_t by (potentially non-deterministically) assigning a *reward* to each of the options in DP_t. We also allow the DP_t and D_t to depend on the history in arbitrary (including adversarial) ways. They need not be drawn from some fixed distribution or the like.

At each time t, an agent chooses from DP_t. The agent receives feedback about its choice. We assume that at the very least the agent is informed of $D_t(c_t)$, its reward for its choice in this round, though other information may also be revealed at time t. 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, our goal is to develop a theory that describes how a rational agent for this problem should learn to choose. The standard theory for rational decision making under uncertainty is Bayesian decision theory (BDT) (Savage, 1954; Jeffrey, 1965; for overviews, see, e.g., Peterson, 2009; Steele & Stefánsson, 2016). 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 aim to develop a theory that can give recommendations for computationally bounded agents. In the following, we give two different kinds of examples to illustrate the role of logical omniscience in Bayesian decision theory and motivate our search for an alternative theory.

3.1. Mere intractability

The first problem with BDT is that in most realistic choice problems, it is hopelessly intractable to follow BDT. Full Bayesian updating or Bayes-optimal decision making are themselves only feasible if the environment is small or highly structured (Savage, 1954, Sections 2.5, 5.5; Cooper, 1990; Chatterjee et al., 2016). Note that even if the agent automatically had a perfectly accurate world model, then determining the optimal choice may require an agent to solve a variety of computationally hard problems, such as the travelling salesman problem, planning in 2-player competitive games (e.g., Even & Tarjan, 1976; Schaefer, 1978), etc. Optimal choice may also rely on whether particular mathematical claims are true, e.g., when assessing the safety of particular cryptographic codes. In all these problems, BDT simply requires the agent to perfectly solve the problem at hand. However, we would like a theory of rational choice that is able to make recommendations for realistic, bounded agents who can only solve such problems approximately.

For illustration, we now give an example of a decision problem in which BDT has little to say to a computationally bounded agent but in which it is especially clear what recommendation we would expect from such a theory. 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. 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 too difficult for any real-world agent.[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 this section with a note on what we can expect

[1] Remote digits of π are a canonical example in the literature on bounded rationality and logical uncertainty (see Savage, 1967, for an early usage). To the knowledge of the authors it is not known 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 (2005).

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 Section 8 of Russell et al.'s (1991; 1993; 1995) work on bounded optimality); or it might tell us at what odds to bet on Goldbach's conjecture with our colleagues. In this paper, we do not provide such a theory and such a theory cannot exist.[2] 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.

3.2. Paradoxes of self-reference, strategic interactions, and counterfactuals

A second problem with BDT and logical omniscience more generally is that the approach of choosing based on a logically omniscient belief system breaks if the values of different options depend on what the agent chooses. As an example, consider the following decision problem, which we will call the Simplified Adversarial Offer (after a decision problem introduced by Oesterheld & Conitzer, 2021). Formally, for any agent c (which, for concreteness, one may imagine is implemented by some computer program), let $\mathrm{SAO}_c = \{a_0, a_1\}$ be the decision problem where a_0 is known to pay off $1/2$ with certainty, and a_1 is known to pay off 1 if (the program implementing) c chooses a_0 in this problem, and 0 otherwise. That is, the reward for a_1 is 1 if and only if c does *not* choose a_1. Now assume for contradiction that c (deterministically) makes the optimal choice 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, c selects a different option than what the belief system predicts. This is a contradiction. Hence, there exists no policy (that can be described in $\mathcal{T}$ and resolved by $\bar{D}$) that complies with standard BDT in this problem. Compare the examples of Oesterheld & Conitzer (2021) and Spencer (2021); also see Demski & Garrabrant (2020, 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. Rubinstein, 1998, Sect. 10.4; Tennenholtz, 2004; van der Hoek et al., 2013; Barasz et al., 2014; Critch, 2016; Oesterheld, 2019). 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, see Binmore (1987) and references therein (cf. Rubinstein, 1998, Ch. 10; Demski & Garrabrant, 2020, Sect. 3.2).

4. The rationality criterion

In this section, we describe our novel rationality requirement, which is the main contribution of this paper.

4.1. Preliminary definitions

An *agent* α chooses at each time t based on past experience one of the available options $\alpha_t^c \in \mathrm{DP}_t$ and moreover provides an *estimate* $\alpha_t^e \in [0,1]$. It is helpful to imagine an agent to be a function that maps any record of past experience – which always must contain at least the rewards obtain from past choices – onto a decision and any given decision problem onto a choice and estimate. However, throughout this paper we will generally only consider the behavior of an agent in a single (though often generic) decision problem sequence. Hence (in line with the multi-armed bandit literature) we leave function application implicit in writing α_t^c and α_t^e. Moreover, we will often identify the agent with the sequence $\bar{\alpha} = (\alpha_t^c, \alpha_t^e)_{t\in\mathbb{N}}$. For example, let $\mathrm{SAO}_{\alpha,t}$ be the Simplified Adversarial Offer for the agent at time t as described in Section 3.2. 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 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. Following Garrabrant et al. (2016), we will often let $\mathbb{H}$ be

[2] For example, Blum's (1967) speedup theorem states, roughly, that there is a decision problem such that for every algorithm solving that decision problem, there exists another, much faster algorithm solving that decision problem. Also, by, e.g., Rice's theorem, it is not even decidable, for a given computational problem, whether it can be solved within some given computational constraints. Also see Hutter (2005, Section 7.1) for some discussion, including a positive result, i.e., an algorithm that is in some sense optimal for all well-defined computational problems.

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, P (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 4.1. For $T \in \mathbb{N}$, we call $\mathcal{L}_T(\bar{\alpha}, \bar{D}) := \sum_{t=1}^{T} \alpha_t^e - D_t(\alpha_t^c)$ the *cumulative overestimation* of an agent $\bar{\alpha}$ on $\bar{D}$.

Definition 4.2. We say $\bar{\alpha}$ for $\bar{D}$ *does not overestimate (on average in the limit)* if $\mathcal{L}_T(\bar{\alpha}, \bar{D})/T \leq 0$ as $T \to \infty$.

In other words, for all $\epsilon > 0$, there should be a time t such that for all $T > t$, $\mathcal{L}_T(\bar{\alpha}, \bar{D})/T \leq \epsilon$. 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 now come to our second requirement, which specifies how the agent $\bar{\alpha}$ relates to the hypotheses in $\mathbb{H}$.

Definition 4.3. 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 4.4. 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 4.5. Let $\bar{h}$ be a hypothesis and $M \subseteq \mathbb{N}$ be a test set of $\bar{\alpha}$ for $\bar{h}$. We call $l_T(\bar{\alpha}, \bar{D}, M, \bar{h}) := \sum_{t \in M_{\leq T}} D_t(h_t^c) - h_t^e$ *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 4.6. 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 $\left(l_T(\bar{\alpha}, \bar{D}, M, \bar{h})\right)_{T \in B}$ 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 4.7. Let $\bar{D}$ be a decision problem sequence and $\bar{\alpha}$ be an agent for $\bar{D}$. Let $\mathbb{H} = \{h_1, h_2, ...\}$ be a set of hypotheses. We say $\bar{\alpha}$ is a *boundedly rational inductive agent (BRIA) for* $\bar{D}$ *covering* $\mathbb{H}$ *with test sets* $M_1, M_2, ...$ 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 Ce 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 and $x_t \in [0, 1]$ with $D_t(x_t) = x_t$. As usual we assume that the 2^t-th binary digits of π are pseudorandom (in a way we will make precise in Section 6.3) 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 $\epsilon > 0$, the hypothesis h_π^ϵ that always recommends a_t^π and estimates $1/2 - \epsilon$. 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 $1/2 - \epsilon$. (Note that if the 2^t-th binary digits of π act like random variables, then this would presumably not be true for $\epsilon = 0$, due to the well-known recurrence (a.k.a. Gambler's ruin) result about the simple symmetric random walk on the line (Pólya, 1921).) With the assumption that $\bar{\alpha}$ covers h_π^ϵ, it follows that for all but finitely many t, $\alpha_t^e \geq 1/2 - \epsilon$. 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 the Simplified Adversarial Offers. In particular at time t, the agent faces $\mathrm{SAO}_{\alpha,t} = \{a_0, a_1\}$, where a_0 pays off $1/2$ with certainty, and a_1 is evaluated to 1 if on the present problem α chooses a_0 and to 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 Section 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 5.1. *Let $\mathbb{H}$ be a computably enumerable set conisting of ($O(g(t))$-)computable hypotheses. (Let $g \in \Omega(\log)$.) Then there exists a BRIA for $\bar{D}$ covering $\mathbb{H}$ that is computable (in $O(g(t)q(t))$, for arbitrarily slow-growing, $O(g(t))$-computable q with $q(t) \to \infty$).*

We here give a sketch of our construction. The idea is that 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 variable. 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.

It can similarly be shown that, for example, a BRIA relative to the class P of hypotheses computable in polynomial time can be computed in arbitrarily close to polynomial time, i.e. in $O(t^{q(t)})$ for arbitrarily slow-growing q with $q(t) \to \infty$.

The next result shows that the BRIAs given by Theorem 5.1 are asymptotically optimal in terms of runtime.

Theorem 5.2. *Let α be a BRIA for $\bar{D}, \mathbb{H}$. 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

6.1. Options with payoff guarantees

Throughout this section, we will show that BRIAs satisfy many desiderata that one might have for rational decision makers. In this section, 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 6.1. *Let $\bar{D}$ be a decision problem sequence and $\bar{\alpha}$ be a BRIA for $\bar{D}$ and the set of e.c. hypotheses. Let $\bar{a}$ be a sequence of terms in $\mathcal{T}$ s.t. for all $t \in \mathbb{N}$, it holds that $a_t \in \mathrm{DP}_t$ and $\alpha_t^c = a_t \implies D_t(a_t) \geq L_t$ 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 $\sum_{t=1}^T D_t(\alpha_t^c)/T \geq \sum_{t=1}^T L_t/T$.*

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 also interpret Theorem 6.1 as providing an immunity to money extraction schemes, which is one of the most widely discussed rationality conditions. 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 the Simplified Adversarial Offer of Section 3.2, a BRIA must walk away with at least $1/2$, which in turn means that it must choose option $a_0 =$ "$1/2$" with frequency 1.

6.2. Options with random payoffs

The following result shows, roughly, that in the limit BRIAs are von Neumann–Morgenstern rational if von Neumann–Morgenstern rational choice is e.c. That is, when choosing between different lotteries whose expected utilities can be computed efficiently, BRIAs converge to choosing the lottery with the highest expected utility. When other, non-lottery options are available, BRIAs must converge to performing at least as well as the best lottery option.

Theorem 6.2. *Let $\bar{D}$ be a decision problem sequence and α be a BRIA for $\bar{D}$. Let $\bar{a}$ be a sequence of terms in $\mathcal{T}$ s.t. $a_t \in \mathrm{DP}_t$ for all $t \in \mathbb{N}$ and the values $D_t(a_t)$ 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 $\sum_{t=1}^T D_t(\alpha_t^c)/T \geq \sum_{t=1}^T \mu_t/T$.*

The proof idea similar to the proof idea for Theorem 6.1. It works by considering hypotheses that recommend a_t and promise $\mu_t - \epsilon$ and noting that the empirical record of such hypotheses goes to $-\infty$ with probability 0.

6.3. Options with algorithmically random payoffs

Theorem 6.2 only tells us something about *true* random variables. But a key goal of our theory is to also be able to assign expected rewards to *pseudo-* or *algorithmically* random sequences, i.e., sequences that are deterministic and potentially even computable, but relevantly unpredictable under computational constraints. We first offer a formal notion of algorithmic randomness.

Definition 6.3. We say a sequence $(D_t(a_t))_{t\in\mathbb{N}}$ 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) given everything revealed by $\bar{D}$ up until time t, we have $\lim_{T\to\infty} \sum_{t\in S_{\leq T}} D_t(a_t) - \mu_t = 0$.

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 6.3 straightforwardly generalizes the standard definition of (unbounded) vMWC randomness (e.g. Downey & Hirschfeldt, 2010, Definition 7.4.1) to non-binary values with means $\bar{\mu}$ other than $1/2$ and computational constraints with outside input (from $\bar{D}$, which could contain an option of the type, "this option pays 0; by the way, the trillionth digit of π is 2"). The notion of vMWC randomness is generally considered quite weak (e.g. Downey & Hirschfeldt, 2010, Sect. 6.2). The most widely studied notion of *bounded* algorithmic randomness is Schnorr bounded randomness (Schnorr, 1971; Ambos-Spies et al., 1997; Wang, 2000; Stull, 2020). An analogous result can be shown w.r.t. this notion.

Theorem 6.4. *Let $\bar{\mu}$ be an e.c. sequence on $[0, 1]$. Let $\bar{D}$ be a decision problem sequence and α be an $O(h(t))$-computable BRIA for $\bar{D}$ covering all e.c. hypotheses. Let $\bar{a}$ be a sequence of terms in $\mathcal{T}$ s.t. $a_t \in \mathrm{DP}_t$ for all $t \in \mathbb{N}$ and the values $D_t(a_t)$ are $O(h(t))$-boundedly vMWC random with means $\bar{\mu}$. Then in the limit as $T \to \infty$, it holds that $\sum_{t=1}^T D_t(\alpha_t^c)/T \geq \sum_{t=1}^T \mu_t/T$.*

7. Boundedly rational inductive agents as a foundation for game theory

We start by briefly giving definitions of the relevant game-theoretic concepts. For a thorough introduction to game theory, see Osborne (2004) or any other textbook on the topic. 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 $u_1, u_2 \colon A_1 \times A_2 \rightarrow [0,1]$. A *(pure) strategy profile* is a pair $(a_1, a_2) \in A_1 \times A_2$. We call (a_1, a_2) a *(pure) Nash equilibrium* if for $i = 1, 2$, $a_i \in \arg\max_{a'_i \in A_i} u_i(a'_i, a_{-i})$, where $-i$ refers to the player other than i. We call the Nash equilibrium strict if both argmaxes are singletons.

Now we imagine that we have two BRIAs $\bar{\alpha}_1, \bar{\alpha}_2$ repeatedly play a particular game against each other. That is, for each $a_1 \in A_1$, $\mathrm{DP}_t^{\alpha_1}$ contains an option that pays $u_1(a_1, a_2)$, where a_2 is the action corresponding to the option selected by α_2 in $\mathrm{DP}_t^{\alpha_2}$. The other agent's decision problem $\mathrm{DP}_t^{\alpha_2}$ is defined analogously. Abusing notation a little, we use $a_i \in A_i$ to represent the available options in $\mathrm{DP}_t^{\alpha_i}$. For instance, we write $\alpha^c_{i,t} = a_1$ to denote that α_i chooses the option from $\mathrm{DP}_t^{\alpha_i}$ that corresponds to $a_i \in A_i$.

How two BRIAs α_1, α_2 play against each other depends on what exactly these BRIAs look like. In particular, if α_1, α_2 always choose the same, then by Theorem 6.1 they will converge to cooperate in the Prisoner's Dilemma, thus deviating from Nash equilibrium play. However, we believe that when two BRIAs do not happen to be correlated in such an extreme way, they will generally only be able to converge to playing a Nash equilibrium (if they converge at all).

We will here prove this under an assumption about non-correlation between $\bar{\alpha}_1$ and $\bar{\alpha}_2$. We start by defining what it means for one player's test set to be uncorrelated in the relevant sense with the other player's choices. We say that a set $M \subseteq \mathbb{N}$ is *weakly uncorrelated* with $\bar{\alpha}_2$ if whenever $\alpha^c_{2,t} = a_2$ with frequency 1 on $t \in \mathbb{N}$ for some $a_2 \in A_2$, $\alpha^c_{2,t} = a_2$ is also true with frequency 1 on $t \in M$.

Next, we define the kind of hypothesis that prevents convergence to non-equilibria, if tested in an uncorrelated way. Let $a_2 \in A_2$ and $a_1^* \in \arg\max_{a_1} u_1(a_1, a_2)$ be a best response to a_2. Also let $\mu = \max\left(\{u_1(a_1, a_2) \mid a_1 \in A_1\} - \{u_1(a_1^*, a_2)\}\right)$ be the utility for Player 1 of playing a second-best response to a_2. Then we call h a safe $a_2 \rightarrow a_1^*$ hypothesis if there is an $\epsilon > 0$ s.t. if $\alpha^c_{2,t} = a_2$ with frequency 1, then $h^c_t = a_1^*$ and $h^e \in [\mu + \epsilon, u_1(a_1^*, a_2) - \epsilon]$ with frequency 1 and otherwise $h^e_t = 0$.

If we assume that there are safe best response hypotheses, convergence is only possible to Nash equilibria.

Theorem 7.1. *Let $\bar{\alpha}_1, \bar{\alpha}_2$ be BRIAs for the decision problem sequences $\bar{D}^{\alpha_1}, \bar{D}^{\alpha_2}$, respectively. Assume that for each player i, and each pair of $a_{-i} \in A_{-i}$ and a best response a_i^* to a_{-i}, there is a safe $a_{-i} \rightarrow a_i^*$ hypothesis that either outpromises $\bar{\alpha}_i$ only finitely many times or whose test set in $\bar{\alpha}_1$ is weakly uncorrelated with $\bar{\alpha}_2$. If $\alpha^c_{1,t}, \alpha^c_{2,t}$ converge to choosing, with frequency 1, the options corresponding to $a_1 \in A_1, a_2 \in A_2$, then (a_1, a_2) is a Nash equilibrium of the underlying game.*

It is not immediately obvious whether safe best response hypotheses will naturally exist (e.g., when the two BRIAs are designed independently without attempts to coordinate). However, we conjecture that convergence to non-equilibria requires deliberately fine tuning the BRIAs to each other.

Finally, we show that for every strict Nash equilibrium, there is a pair of BRIAs that converge to that Nash equilibrium.

Theorem 7.2. *For each game (A_1, A_2, u_1, u_2) and strict Nash equilibrium $(a_1, a_2) \in A_1 \times A_2$, there is a pair of randomizing agents $\bar{\alpha}_1$ and $\bar{\alpha}_2$ that are BRIAs with probability 1 relative to any (countable) set of hypotheses $\mathbb{H}$ and that converge to playing (a_1, a_2) with probability 1.*

8. Related work

Contextual stochastic multi-armed bandits As noted in Section 2, our problem is a contextual multi-armed bandit problem as considered in statistical learning theory. However, papers in this literature generally avoid the possibility of an environment that can refer to the agent (as in the Adversarial Offer or strategic interactions). For example, Yang & Zhu (2002, Assumption A in Section 5) and Agarwal et al. (2012, Assumption 1 in Section 2) assume that the agent's models can converge to being accurate. These assumptions allow a much simpler rationality requirement, namely some kind of convergence to optimal behavior (cf. Section 6.1).

Adversarial multi-armed bandits with expert advice Another closely related literature is that on multi-armed bandit problems with expert advice (Auer et al., 2001, Section 7; Lattimore & Szepesvari, 2017, Chapter 18). This literature generally allows adversarial problems. Like this paper, it addresses this problem by making the optimality goal relative to some set of hypotheses. However, its optimality criterion is quite 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 $\bar{D}$ (see, e.g, the Exp4 algorithm of Auer et al., 2001, Section 7).

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 B. At the same time, we conjecture that learners with low regret relative to a set of hypotheses $\mathbb{H}$ satisfy a version of the BRIA criterion; see Appendix C for a preliminary result.

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 (Nozick, 1969; Ahmed, 2014).

Most of this literature has focused on discussing relatively simple cases (similar to the Simplified Adversarial Offer). In these cases, BRIAs generally side with what has been called evidential decision theory. For example, by Theorem 6.1, 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 Section 6.2). 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 Section 3, especially Section 3.1, 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., Selten, 1990, 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. Conlisk, 1996). For instance, for most humans even 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 Section 3.2 (Binmore, 1987; Rubinstein, 1998, Ch. 10).

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 will discuss a few approaches associated with the bounded rationality literature that take a similar approach as 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* (Russell & Wefald, 1991, Sect. 1.4; Russell et al., 1993; Russell & Subramanian, 1995). 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 Section 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. Demski & Garrabrant, 2020, 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 Section 3.2 in problematic ways.

Garrabrant inductors As noted earlier, the present approach to dealing with computational constraints is inspired by the work of Garrabrant et al. (2016), who address the problem of assigning probabilities under computational constraints. In particular, their approach allows assigning probabilities to logical statements (such as: "the 10^{10}-th digit of π is 0"). 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 (2017). 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 B). We conjecture that, like regret minimizers, 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. (2016)? 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?

References

Agarwal, A., Dudík, M., Kale, S., Langford, J., and Schapire, R. E. Contextual bandit learning with predictable rewards. In *Appearing in Proceedings of the 15th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 22 of *Proceedings of Machine Learning Research*, pp. 19–26. La Palma, Canary Islands, 4 2012. URL http://proceedings.mlr.press/v22/agarwal12/agarwal12.pdf.

Ahmed, A. *Evidence, Decision and Causality*. Cambridge University Press, 2014.

Ambos-Spies, K., Terwijn, S. A., and Xizhong, Z. *Theoretical Computer Science*, 172(1–2):195–207, 2 1997. doi: 10.1016/S0304-3975(95)00260-X.

Arntzenius, F. No regrets, or: Edith piaf revamps decision theory. *Erkenntnis*, 68:277–297, 2008. doi: 10.1007/s10670-007-9084-8.

Auer, P., Cesa-Bianchi, N., Freund, Y., and Shapire, R. E. The non-stochastic multi-armed bandit problem, 11 2001. URL https://cseweb.ucsd.edu/~yfreund/papers/bandits.pdf.

Barasz, M., Christiano, P., Fallenstein, B., Herreshoff, M., LaVictoire, P., and Yudkowsky, E. Robust cooperation in the prisoner's dilemma: Program equilibrium via provability logic, 1 2014. URL https://arxiv.org/abs/1401.5577.

Bell, J., Linsefors, L., Oesterheld, C., and Skalse, J. Reinforcement learning in Newcomblike environments. In *Advances in Neural Information Processing Systems 34 proceedings (NeurIPS 2021)*. 2021. URL https://proceedings.neurips.cc/paper/2021/file/b9ed18a301c9f3d183938c451fa183df-Paper.pdf.

Binmore, K. Modeling rational players: Part I. *Economics & Philosophy*, 3(2):179–214, 10 1987.

Blum, M. A machine-independent theory of the complexity of recursive functions. *Journal of the Association for Computing Machinery*, 14(2):322–336, 4 1967.

Chatterjee, K., Chmelík, M., and Tracol, M. What is decidable about partially observable Markov decision processes with ω-regular objectives. *Journal of Computer and System Sciences*, 82(5):878–911, 8 2016.

Conlisk, J. Why bounded rationality? *Journal of economic literature*, 34(2):669–700, 1996.

Cooper, G. F. The computational complexity of probabilistic inference using bayesian belief networks. *Artificial Intelligence*, 42(2-3):393–405, 3 1990. doi: 10.1016/0004-3702(90)90060-D.

Critch, A. Parametric bounded Löb's theorem and robust cooperation of bounded agents, 2016. URL https://arxiv.org/abs/1602.04184.

Demski, A. and Garrabrant, S. Embedded agency, 2020. URL https://arxiv.org/pdf/1902.09469.pdf.

Downey, R. G. and Hirschfeldt, D. R. *Algorithmic randomness and complexity*. Springer, 2010.

Even, S. and Tarjan, R. E. A combinatorial problem which is complete in polynomial space. *Journal of the ACM*, 23 (4):710–719, 10 1976. doi: 10.1145/321978.321989.

Garrabrant, S. Two major obstacles for logical inductor decision theory, 4 2017. URL https://www.alignmentforum.org/posts/5bd75cc58225bf06703753d4/two-major-obstacles-for-logical-inductor-decision

Garrabrant, S. Temporal inference with finite factored sets, 2021. URL https://arxiv.org/pdf/2109.11513.pdf.

Garrabrant, S., Benson-Tilsen, T., Critch, A., Soares, N., and Taylor, J. Logical induction, 2016. URL https://intelligence.org/files/LogicalInduction.pdf. A short version is available at https://intelligence.org/files/LogicalInductionAbridged.pdf; another short version was published in TARK '17, see https://arxiv.org/abs/1707.08747.

Harper, W. L. Mixed strategies and ratifiability in causal decision theory. *Erkenntnis*, 24(1):25–36, 1 1986. URL https://www.jstor.org/stable/20006545.

Hutter, M. *Universal Artificial Intelligence. Sequential Decision Based on Algorithmic Probability*. Texts in Theoretical Computer Science. Springer, 2005.

Jeffrey, R. C. *The Logic of Decision*. McGraw-Hill, New York, 1965.

Lattimore, T. and Szepesvari, C. Bandit algorithms. 2017. URL https://tor-lattimore.com/downloads/book/book.pdf.

Levinstein, B. A. and Soares, N. Cheating death in damascus. *The Journal of Philosophy*, 117(5):237–266, 5 2020. doi: 10.5840/jphil2020117516.

Marsaglia, G. On the randomness of pi and other decimal expansions. *InterStat*, 10 2005. URL https://web.archive.org/web/20210412122444/http://interstat.statjournals.net/YEAR/2005/articles/0510005.pdf.

Nozick, R. Newcomb's problem and two principles of choice. In et al., N. R. (ed.), *Essays in Honor of Carl G. Hempel*, pp. 114–146. Springer, 1969. URL http://faculty.arts.ubc.ca/rjohns/nozick_newcomb.pdf.

Oesterheld, C. Robust program equilibrium. *Theory and Decision*, 86(1):143–159, 2 2019.

Oesterheld, C. and Conitzer, V. Extracting money from causal decision theorists. *The Philosophical Quarterly*, 2021. doi: https://doi.org/10.1093/pq/pqaa086.

Osborne, M. J. *An Introduction to Game Theory*. Oxford University Press, 2004.

Peterson, M. *An Introduction to Decision Theory*. Cambridge University Press, 2009.

Pólya, G. Über eine Aufgabe der Wahrscheinlichkeitsrechnung betreffend die Irrfahrt im Straßennetz. *Mathematische Annalen*, 84(1):149–160, 1921.

Richter, R. Rationality revisited. *Australasian Journal of Philosophy*, 62(4):392–403, 1984. doi: 10.1080/00048408412341601.

Rubinstein, A. *Modeling Bounded Rationality*. Zeuthen Lecture Book Series. The MIT Press, 1998.

Russell, S. and Wefald, E. *Do the Right Thing – Studies in Limited Rationality*. The MIT Press, 1991.

Russell, S. J. and Subramanian, D. Provably bounded-optimal agents. *Journal of Artificial Intelligence Research*, 2:575–609, 5 1995. doi: 10.1613/jair.133.

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

Savage, L. J. *The Foundations of Statistics*. John Wiley and Sons, New York, 1954.

Savage, L. J. Difficulties in the theory of personal probability. *Philosophy of Science*, 34(4):305–310, 12 1967. URL https://www.jstor.org/stable/186119.

Schaefer, T. J. On the complexity of some two-person perfect-information games. *Journal of Computer and System Sciences*, 16(2):185–225, 4 1978. doi: 10.1016/0022-0000(78)90045-4.

Schnorr, C. P. *Zufälligkeit und Wahrscheinlichkeit. Eine algorithmische Begründung der Wahrscheinlichkeitstheorie*, volume 218 of *Lecture Notes in Mathematics*. Springer, 1971.

Selten, R. Bounded rationality. *Journal of Institutional and Theoretical Economics*, 146(4):649–658, 12 1990.

Skyrms, B. Deliberational equilibria. *Topoi*, 5:59–67, 3 1986. doi: 10.1007/BF00137830.

Spencer, J. An argument against causal decision theory. *Analysis*, 81(1):52–61, 2021.

Steele, K. and Stefánsson, H. O. Decision theory. In Zalta, E. N. (ed.), *The Stanford Encyclopedia of Philosophy*. Winter 2016 edition, 2016. URL https://plato.stanford.edu/archives/win2016/entries/decision-theory/.

Stull, D. M. Resource bounded randomness and its applications. In *Algorithmic Randomness – Progress and Prospects*, Lecture Notes in Logic, pp. 301–348. Cambridge University Press, 2020.

Sutton, R. S. and Barto, A. G. *Reinforcement Learning: An Introduction*. MIT Press, 1998.

Tennenholtz, M. Program equilibrium. *Games and Economic Behavior*, 49(2):363–373, Nov 2004.

Thorndike, E. L. *Animal Intelligence – Experimental Studies*. The Macmillan Company, New York, 6 1911.

van der Hoek, W., Witteveen, C., and Wooldridge, M. Program equilibrium – a program reasoning approach. *International Journal of Game Theory*, 42:639–671, Aug 2013.

Wang, Y. Resource bounded randomness and computational complexity. *Theoretical Computer Science*, 237(1–2): 33–55, 4 2000.

Weirich, P. Causal decision theory. In *The Stanford Encyclopedia of Philosophy*. Spring 2016 edition, 2016.

Yang, Y. and Zhu, D. Randomized allocation with nonparametric estimation for a multi-armed bandit problem with covariates. *The Annals of Statistics*, 30(1):100–121, 2002.

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.

Lemma A.1. *Let $\bar{h}$ be a hypothesis and $N \subseteq \mathbb{N}$ s.t. $t \in N$ implies $h^e_t = 0$. 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, it is

$$\begin{aligned} l_T(\bar{\alpha}, \bar{D}, M, \bar{h}) &= \sum_{t \in M_{\leq T}} D_t(h^c_t) - h^e_t \\ &= \sum_{t \in M_{\leq T} - N} D_t(h^c_t) - h^e_t + \sum_{t \in M_{\leq T} \cap N} D_t(h^c_t) - h^e_t \\ &= \sum_{t \in M_{\leq T} - N} D_t(h^c_t) - h^e_t + \sum_{t \in M_{\leq T} \cap N} D_t(h^c_t) \\ &\geq \sum_{t \in M_{\leq T} - N} D_t(h^c_t) - h^e_t \\ &= l_t(\bar{\alpha}, \bar{D}, M - N, \bar{h}). \end{aligned}$$

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

A.2. Proof of Theorem 5.1

Theorem 5.1. *Let $\mathbb{H}$ be a computably enumerable set conisting of ($O(g(t))$-)computable hypotheses. (Let $g \in \Omega(\log)$.) Then there exists a BRIA for $\bar{D}$ covering $\mathbb{H}$ that is computable (in $O(g(t)q(t))$, for arbitrarily slow-growing, $O(g(t))$-computable q with $q(t) \to \infty$).*

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 average 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[3]) auction for the present decision problem among all hypotheses. That is, we determine a winning hypothesis

$$i^*_t \in \arg\max_{i \in \mathbb{N}} \min(h^e_{i,t}, w_t(i)) \tag{2}$$

[3] 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.

with arbitrary tie breaking. Intuitively, each hypothesis h_i bids $h^e_{i,t}$, 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 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 $\alpha_t \coloneqq (h^c_{i^*_t,t}, e^*_t)$.

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 $w_{t+1}(i^*_t) \leftarrow w_t(i^*_t)+A(t,i^*_t)+D_t(h^c_{i^*_t,t})-e^*_t$. That is, the highest-bidding hypothesis receives the allowance and the reward obtained after following its recommendation ($D_t(h^c_{i^*_t,t})$), 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

$$\sum_{i\in B^+_T} w_T(i) = \sum_{i\in B^+_T}\sum_{n=1}^{T} A(n,i) + \sum_{t=1}^{T} D_t(\alpha^c_t) - \alpha^e_t.$$

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:

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

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 4.4), 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 $\min(h^e_{i,t}, w_t(i)) = h^e_{i,t}$. But it is infinitely often $h^e_{i,t} > \alpha^e_t$. 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

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

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

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

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 $\mathbb{H}$, $\bar{D}$ 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 5.2

Theorem 5.2. *Let α be a BRIA for $\bar{D}, \mathbb{H}$. 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 $\alpha^e_t < 1$, promises 1 and recommends an option other than α^c_t; 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 6.1

Theorem 6.1. *Let $\bar{D}$ be a decision problem sequence and $\bar{\alpha}$ be a BRIA for $\bar{D}$ and the set of e.c. hypotheses. Let $\bar{a}$ be a sequence of terms in $\mathcal{T}$ s.t. for all $t \in \mathbb{N}$, it holds that $a_t \in \mathrm{DP}_t$ and $\alpha^c_t = a_t \implies D_t(a_t) \geq L_t$ 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 $\sum_{t=1}^{T} D_t(\alpha^c_t)/T \geq \sum_{t=1}^{T} L_t/T$.*

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

We prove this new claim by proving a contrapositive. In particular, we assume that $\alpha^e_t < L_t$ 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 $D_t(a_t) \geq L_t$. It follows that for all T,

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

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

A.5. Proof of Theorem 6.2

Theorem 6.2. *Let $\bar{D}$ be a decision problem sequence and α be a BRIA for $\bar{D}$. Let $\bar{a}$ be a sequence of terms in $\mathcal{T}$ s.t. $a_t \in \mathrm{DP}_t$ for all $t \in \mathbb{N}$ and the values $D_t(a_t)$ 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 $\sum_{t=1}^{T} D_t(\alpha_t^c)/T \geq \sum_{t=1}^{T} \mu_t/T$.*

Proof. We need only show that with probability 1 for all $\epsilon > 0$ it holds that for all but finitely many times t that $\alpha_t^e \geq \mu_t - \epsilon$. 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 $\epsilon > 0$ s.t. with some positive probability $p > 0$ we infinitely often have that $\alpha_t^e < \mu_t - \epsilon$, then $\bar{\alpha}$ is with positive probability not a BRIA.

Consider the hypothesis $\bar{h}_{a,\epsilon}$ that at each time t promises $\max(\mu_t - \epsilon, 0)$ and recommends a_t. Since with probability p, $\bar{h}_{a,\epsilon}$ 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 $\bar{h}_{a,\epsilon}$ is tested infinitely often), let the test set be some infinite set $M \subseteq \mathbb{N}$. (Note that M may depend on $\bar{D}$ and inherit its stochasticity. This will not matter for the following, though.) For simplicity, let M be the empty set if $\bar{h}_{a,\epsilon}$ does not outpromise α infinitely often. By Lemma A.1, we can assume WLOG that for all $t \in M$, $h_{a,\epsilon}^e = \mu_t - \epsilon$. Now notice that

$$\frac{1}{|M_{i,\leq T}|} l_T(\alpha, \bar{D}, M_i, \bar{h}_i) = \frac{1}{|M_{i,\leq T}|} \sum_{t \in M_{i,\leq T}} D_t(h_{a,\epsilon,t}^c) - h_{a,\epsilon,t}^e = \frac{1}{|M_{i,\leq T}|} \sum_{t \in M_{i,\leq T}} D_t(a_t) - (\mu_t - \epsilon).$$

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 $\sum_{t \in M_{i,\leq T}} D_t(a_t) - (\mu_t - \epsilon) \underset{\text{w.p. } 1}{\to} \epsilon$ as $T \to \infty$ by the law of large numbers. We have thus shown that with positive probability (p) $\bar{h}_{a,\epsilon}$ outpromises $\bar{\alpha}$ infinitely often while $\bar{h}_{a,\epsilon}$'s record $l_T(\alpha, \bar{D}, M_i, \bar{h}_i)$ is positive in all but finitely many rounds. Thus, in this positive-probability event $\bar{\alpha}$'s infinitely many rejections of $\bar{h}_{a,\epsilon}$ violates the coverage criterion. □

A.6. Proof of Theorem 6.4

Definition A.2. We say a sequence $(D_t(a_t)))_{t \in \mathbb{N}}$ 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) given everything revealed by $\bar{D}$ up until time t, we have $\lim_{T \to \infty} \sum_{t \in S_{\leq T}} D_t(a_t) - \mu_t = 0$.

Theorem 6.4. *Let $\bar{\mu}$ be an e.c. sequence on $[0, 1]$. Let $\bar{D}$ be a decision problem sequence and α be an $O(h(t))$-computable BRIA for $\bar{D}$ covering all e.c. hypotheses. Let $\bar{a}$ be a sequence of terms in $\mathcal{T}$ s.t. $a_t \in \mathrm{DP}_t$ for all $t \in \mathbb{N}$ and the values $D_t(a_t)$ are $O(h(t))$-boundedly vMWC random with means $\bar{\mu}$. Then in the limit as $T \to \infty$, it holds that $\sum_{t=1}^{T} D_t(\alpha_t^c)/T \geq \sum_{t=1}^{T} \mu_t/T$.*

Proof. We prove the theorem by proving that for all $\epsilon > 0$, $\alpha_t^e \geq \mu_t - \epsilon$ 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 $\bar{h}_{a,\epsilon}$ that at each time t promises $\max(\mu_t - \epsilon, 0)$ and recommends a_t. Since $\bar{h}_{a,\epsilon}$ infinitely often outpromises $\bar{\alpha}$, it must tested infinitely often. Let the test set be some infinite set $M \subseteq \mathbb{N}$. By Lemma A.1, we can assume WLOG that for all $t \in M$, $h_{a,\epsilon}^e = \mu_t - \epsilon$.

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

$$\frac{1}{|M_{i,\leq T}|} l_T(\alpha, \bar{D}, M_i, \bar{h}_i) = \frac{1}{|M_{i,\leq T}|} \sum_{t \in M_{i,\leq T}} D_t(a_t) - (\mu_t - \epsilon) \underset{\text{w.p. } 1}{\rightarrow} \epsilon \text{ as } T \rightarrow \infty,$$

where the final step is by the fact that $(D_t(a_t))_t$ is vMWC random with means $\bar{\mu}$. Hence, with probability 1, $\bar{h}_{a,\epsilon}$'s record $l_T(\alpha, \bar{D}, M_i, \bar{h}_i)$ is positive in all but finitely many rounds. Thus, $\bar{\alpha}$'s infinitely many rejections of $\bar{h}_{a,\epsilon}$ violates the coverage criterion. □

A.7. Proof of Theorem 7.1

Theorem 7.1. *Let $\bar{\alpha}_1, \bar{\alpha}_2$ be BRIAs for the decision problem sequences $\bar{D}^{\alpha_1}, \bar{D}^{\alpha_2}$, respectively. Assume that for each player i, and each pair of $a_{-i} \in A_{-i}$ and a best response a_i^* to a_{-i}, there is a safe $a_{-i} \rightarrow a_i^*$ hypothesis that either outpromises $\bar{\alpha}_i$ only finitely many times or whose test set in $\bar{\alpha}_1$ is weakly uncorrelated with $\bar{\alpha}_2$. If $\alpha^c_{1,t}, \alpha^c_{2,t}$ converge to choosing, with frequency 1, the options corresponding to $a_1 \in A_1, a_2 \in A_2$, then (a_1, a_2) is a Nash equilibrium of the underlying game.*

Proof. We prove this by contradiction. That is, we assume that $\bar{\alpha}_1, \bar{\alpha}_2$ converge to choosing some non-NE (a_1, a_2) with frequency 1, and then show a contradiction to the assumption that $\bar{\alpha}_1, \bar{\alpha}_2$ are BRIAs.

WLOG let there be a best response $a_1^* \in A_1$ s.t. $u_1(a_1^*, a_2) > u_1(a_1, a_2)$. Then consider a safe, weakly uncorrelated $a_2 \rightarrow a_1^*$ hypothesis $\bar{h}_i$.

First we show that $\bar{\alpha}_1$ infinitely often rejects h_i. The no overestimation criterion applied to $\bar{\alpha}_1$ states that $\sum_{t=1}^T (\alpha^e_{1,t} - D_t(\alpha^c_{1,t}))/T \leq 0$ as $T \rightarrow 0$. Now by the assumption that (a_1, a_2) is played with limit frequency 1, $\sum_{t=1}^T D_t(\alpha^c_{1,t})/T \rightarrow u_1(a_1, a_2)$ as $T \rightarrow 0$. Hence, $\sum_{t=1}^T \alpha^e_{1,t}/T \leq u_1(a_1, a_2)$ as $T \rightarrow 0$. It follows in particular that for all $\epsilon > 0$ with positive limit frequency among $t \in \mathbb{N}$, we have that $\alpha^e_{1,t} < u_1(a_1, a_2) + \epsilon$. Because α_2 plays a_2 with limit frequency 1, h_i (by definition of a safe $a_2 \rightarrow a_1^*$ hypothesis) therefore promises above $u_1(a_1, a_2) + \epsilon$ for some ϵ with limit frequency 1 and therefore infinitely often outpromises $\bar{\alpha}$.

Hence there must be an infinite test set M for h_i. As usual, we will assume WLOG (by Lemma A.1) that M includes only rounds in which $\bar{h}_i$ submits non-zero promises. Now consider the average empirical record

$$\frac{1}{|M_{\leq T}|} l_T(\bar{\alpha}_1, \bar{D}^{\alpha_1}, M, \bar{h}) = \frac{1}{|M_{\leq T}|} \sum_{t \in M_{\leq T}} D_t(h^c_{i,t}) - \frac{1}{|M_{\leq T}|} \sum_{t \in M_{\leq T}} h^e_{i,t}. \tag{3}$$

By assumption, α_2 chooses a_2 with limit frequency 1. From this it follows that $\bar{h}_i$ recommends a_1^* with limit frequency 1. By the assumption about weakly uncorrelated testing of $\bar{h}_i$, it also follows that α_2 chooses a_2 with limit frequency 1 on M. From this, it is easy to show that first average converges to $u_1(a_1^*, a_2)$. Since $h^e_{i,t} \leq u_1(a_1^*, a_2) - \epsilon$ for some (constant) $\epsilon > 0$, the second is always less then $u_1(a_1^*, a_2) - \epsilon$. It follows that $l_T(\bar{\alpha}_1, \bar{D}^{\alpha_1}, M, \bar{h})/|M_{\leq T}| \geq \epsilon$ in the limit and therefore also $l_T(\bar{\alpha}_1, \bar{D}^{\alpha_1}, M, \bar{h}) \rightarrow +\infty$, violating the coverage criterion. □

A.8. Proof of Theorem 7.2

Theorem 7.2. *For each game (A_1, A_2, u_1, u_2) and strict Nash equilibrium $(a_1, a_2) \in A_1 \times A_2$, there is a pair of randomizing agents $\bar{\alpha}_1$ and $\bar{\alpha}_2$ that are BRIAs with probability 1 relative to any (countable) set of hypotheses $\mathbb{H}$ and that converge to playing (a_1, a_2) with probability 1.*

Proof. We construct the BRIAs as follows. Basically we use the same construction as that in Appendix A.2 for the special case of $\mathbb{S} = \{\mathbb{N}\}$. However, we add onto this that in every round with some fixed probability $p \in (0, 1)$, the market chooses the equilibrium action a_i regardless of the highest-bidding hypothesis' recommendation. The estimate in these rounds is nonetheless that of the highest-bidding hypothesis. In these replacement rounds, no hypothesis is tested and therefore no hypothesis spends allowance money. The constant p is picked in such a way that it is ensured that the unique best response to this market is always the other player's equilibrium action a_{-i}.

We have to show that agents constructed in this way are indeed BRIAs with probability 1 and that they almost surely converge to playing the given equilibrium (a_1, a_2) with frequency 1.

No overestimation: We have to show that per-round overestimation goes to 0 with probability 1. Let R be the (i.i.d. randomly selected) set of rounds in which a_i is played by "replacement" without any testing. It is

$$\frac{1}{T}\mathcal{L}_T(\bar{\alpha}_1, \bar{D}, \mathbb{N}) = \frac{1}{T}\sum_{T=1}^{T} \alpha^e_{1,t} - D_t(\alpha^c_{1,t}) \tag{4}$$

$$= \frac{1}{T}\sum_{t\in R_{\leq T}} h^e_{i^*_t,t} - D_t(a_i) + \frac{1}{T}\sum_{t\in\{1,\dots,T\}-R} h^e_{i^*_t,t} - D_t(h^c_{i^*_t,t}) \tag{5}$$

$$\leq \frac{1}{T}\sum_{t\in R_{\leq T}} h^e_{i^*_t,t} - D_t(h^c_{i^*_t,t}) + \frac{1}{T}\sum_{t\in\{1,\dots,T\}-R} h^e_{i^*_t,t} - D_t(h^c_{i^*_t,t}) \text{ w.p. 1 as } T\rightarrow\infty. \tag{6}$$

The last step is due to the fact that by construction, a_i is always optimal in expectation. Now, the first of the two summands in the last line can be shown to approach 0 by the same argument that we used in the proof of non overestimation of our BRIA algorithm in Appendix A.2: the cumulative loss is bound by allowance distributed (plus the negligible initial wealth) and per-round-allowance goes to 0. But now notice that the second and first sums are the same, except that they are over complementary sets. However, since R is randomly sampled, the terms must approach each other on average, as follows:

$$\frac{1}{Tp}\sum_{t\in R_{\leq T}} h^e_{i^*_t,t} - D_t(h^c_{i^*_t,t}) - \frac{1}{T(1-p)}\sum_{t\in\{1,\dots,T\}-R} h^e_{i^*_t,t} - D_t(h^c_{i^*_t,t}) \rightarrow 0 \text{ w.p. 1 as } T\rightarrow\infty.$$

Hence, because the latter sum is bound by allowance, the former sum is in the limit almost surely bounded by allowance times $1/p$. We conclude that both summands in line 6 approach 0 and therefore that the no overestimation criterion is satisfied.

Hypothesis coverage: The low *relative* loss property can be shown in the same way as in the proof of Theorem 5.1 in Appendix A.2: whenever a hypothesis strictly outpromises α_1, it must by construction of $\bar{\alpha}_1$ have insufficient wealth. This in turn implies by how wealth in the construction works that the hypothesis must have empirically underperformed its estimates.

Convergence to (a_1, a_2): Finally, we need to prove that these BRIAs indeed almost surely converge to playing (a_1, a_2) with frequency 1, i.e., that each player plays a_i with frequency 1 rather than just with frequency p. This can be shown by essentially the same argument as the proof of Theorem 6.2 in Appendix A.5. By choice of p, recommending a_i guarantees an expected value that is greater than that of any other action. □

B. 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 Section 3.2 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 Section 8). In the literature on such Newcomb-like problems (see Section 8), an idea closely related to regret minimization has been discussed under the name ratificationism (see Weirich, 2016, for an introduction and overview). Ratificationism similarly uses distributions over actions (see, e.g., the formal description by Bell et al., 2021), though often these are not meant to arise from randomization (e.g. Arntzenius, 2008).

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.[4]

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 (Nozick, 1969); see Section 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 Section 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.

C. Some regret minimizers satisfy a generalized BRIA criterion

We here show that some regret minimizers satisfy a slightly generalized version of the BRIA criterion. We first have to give a formal definition of regret. Since the literature on adversarial bandit problems with expert advice does not consider experts who submit estimates in the way that our hypotheses do, we cannot use an existing definition and will instead make up our own. For simplicity, we will only consider the case $\mathbb{S} = \mathbb{N}$.

Let $\bar{D}$ be a decision process, $\bar{\alpha}$ be an agent and $\mathbb{H} = \{h_1, h_2, ...\}$ be a set of hypotheses. For simplicity, let $\mathbb{H}$ be finite. For each $h_i \in \mathbb{H}$, let $B_i := \{t \in \mathbb{N} \mid h^e_{i,t} > \alpha^e_t\}$ be the set of rounds in which h_i outpromises α. We define the average per-round regret of the learner to hypothesis h_i up to time T as

$$\mathrm{REGRET}_{m,T} = \mathbb{E}\left[\frac{1}{|B_{m,\leq T}|}\sum_{t\in B_{m,\leq T}} D_t(h^c_{i,t}) - \alpha^e_t\right].$$

As before, the bidding mechanisms means that hypotheses can specialize on specific types of decisions.[5] As is common in the adversarial bandit problem literature, we will be interested in learning algorithms that guarantee average regret to go to zero as $|B_{m,\leq T}| \to \infty$.

Regret is somewhat analogous to the cumulative empirical record on the test set. As with the coverage condition, low regret can be achieved trivially by setting $\alpha^e = 1$. Thus, if we replace the coverage criterion with a sublinear-regret requirement, we have to keep the no overestimation criterion.

[4] For brief discussions of this and closely related concerns in the literature on Newcomb-like problems, see Richter (1984), Harper (1986), Skyrms (1986), Arntzenius (2008, Section 9), Levinstein & Soares (2020), and Oesterheld & Conitzer (2021, Section IV.1).

[5] Note that we subtract the agent's *estimates*, not the utility that $\bar{\alpha}$ in fact achieves. This is important. Otherwise, the learner can set $\alpha^e = 0$ even in rounds in which $D_t(\alpha^c_t)$ is (expected to be) high, thus circumventing the expert's bidding mechanism.

Still, there are alternative definitions that also work. For example, one might count regret only in rounds in which α and h_i differ in their recommendations.

Conjecture C.1. Let $\bar{D}$ be a decision process where $|\mathrm{DP}_t|$ is bounded for all $t \in \mathrm{N}$. With access to an independent source of randomization, and given access to the outputs of all hypotheses in $\mathbb{H}$, we can compute $\bar{\alpha}$ that does not overestimate on $\mathbb{N}$ s.t. for all hypotheses h_i, $\mathrm{REGRET}_{i,T} \to 0$ with probability 1 if $|B_{i,\leq T}| \to \infty$.

As noted elsewhere, without independent randomization it is clear that such an $\bar{\alpha}$ cannot be designed. Even with independent randomization, it is not obvious whether the conjecture holds. However, similar results in the literature on adversarial bandit problems with expert advice lead us to believe that it does. That said, we have not been able to prove the conjecture by using simply the results from that literature.

Theorem C.2. *Let $\bar{\alpha}$ be an independently randomized agent that does not overestimate on $\mathbb{N}$ and ensures sublinear regret with probability 1 relative to all hypotheses in some finite set $\mathbb{H} = \{h_i\}_i$. Further assume that for all hypotheses h_i, $P(\alpha^c_t{=}h^c_{i,t}) \in \omega(1/t)$ among $t \in B_i$. Then we can compute based on α a new agent $\tilde{\alpha}$ that does not overestimate and that satisfies for each hypothesis h_i that is infinitely often rejected,*

$$\sum_{t \in B_{i,\leq T}} \frac{\mathbb{1}[\alpha^c_t{=}h^c_{i,t}]}{P(\alpha^c_t{=}h^c_{i,t})} (D_t(h^c_{i,t}) - h^e_{i,t}) \to -\infty \tag{7}$$

among T at which $\tilde{\alpha}$ rejects h_i.

Notice that the left-hand side of line 7 is a weighted version of the cumulative empirical record on the set $\{t \in B_{i,\leq T} \mid \alpha^c_t{=}h^c_{i,t}\}$.

The proof combines one key idea from the literature on adversarial multi-armed bandits – importance-weighted estimation – and one from this paper – the decision auction construction (Appendix A.2).

Proof. For $t \in B_i$, define

$$\hat{\mathrm{R}}_{i,t} = \frac{\mathbb{1}[\alpha^c_t{=}h^c_{i,t}]}{P(\alpha^c_t{=}h^c_{i,t})} (D_t(h^c_{i,t}) - \alpha^e_t),$$

where we assume $P(\alpha^c_t{=}h^c_{i,t}) > 0$. As usual we then have that $\mathbb{E}\left[\hat{\mathrm{R}}_{i,t}\right] = D_t(h^c_{i,t}) - \alpha^e_t$. For $t \notin B_i$, define $\hat{\mathrm{R}}_{i,t} = 0$. Hence, $\hat{\mathrm{R}}_{i,t}$ can be used as an unbiased estimator of the regret in a single round. Further, $\mathrm{Var}(\hat{\mathrm{R}}_{m,t}) \in o(t)$, and thus $\sum_{t=1}^{T} \mathrm{Var}(\hat{\mathrm{R}}_{m,t}) \in o(T^2)$. By Kolmogorov's strong law of large numbers,

$$\frac{1}{T} \sum_{t \in B_{m,\leq T}} \hat{\mathrm{R}}_{m,t} - \frac{1}{T} \sum_{t \in B_{m,\leq T}} D_t(h^c_{i,t}) - \alpha^e_t \frac{1}{T} \sum_{t=1}^{T} \hat{\mathrm{R}}_{m,t} - \frac{1}{T} \sum_{t=1}^{T} D_t(h^c_{i,t}) - \alpha^e_t 0 \text{ as } T \to \infty$$

In other terms,

$$\sum_{t \in B_{m,\leq T}} \hat{\mathrm{R}}_{m,t} - \sum_{t \in B_{m,\leq T}} D_t(h^c_{i,t}) - \alpha^e_t$$

is sublinear.

We now construct new estimates. Fix a non-decreasing, sublinear function $\mathrm{CA}\colon \mathbb{N} \to \mathbb{R}$ with $\mathrm{CA}(T) \to \infty$. (These are cumulative versions of the allowance functions from the construction in Appendix A.2.) Next, we define

$$\mathcal{L}_{i,T} \coloneqq \sum_{t \in M_{i,\leq T}} \frac{\mathbb{1}[\alpha^c_t{=}h^c_{i,t}]}{P(\alpha^c_t{=}h^c_{i,t})} (D_t(h^c_{i,t}) - h^e_{i,t}) \sum_{t \in B_{i,\leq T} - M_i} \frac{\mathbb{1}[\alpha^c_t{=}h^c_{i,t}]}{P(\alpha^c_t{=}h^c_{i,t})} (D_t(h^c_{i,t}) - \alpha^e_t),$$

where $M_i \subseteq B_i$ will be defined in a second. Define $w_T(i) = \mathrm{CA}(T) + \mathcal{L}_{i,T}$. Now at each time t, we define our new estimate as

$$\tilde{\alpha}^e_t = \max(\alpha^e_t, \max_{i : w_{t-1}(i) \geq 0} h^e_{i,t}). \tag{8}$$

Finally, let M_i be the set of rounds in which i is the maximizer in Eq. 8 through the outer max.

We now need to show two things: That cumulative overestimation is still sublinear even for the new increased $\tilde{\alpha}^e_t$ and that the claimed variant of the hypothesis coverage criterion is satisfied.

We start with hypothesis coverage. First notice that because $M_i \subseteq B_i$ and for $t \in B_i$, $h^e_{i,t} > \alpha^e_t$, we get that

$$w_T(i) \geq \mathrm{CA}(T) + \sum_{t \in B_{m,\leq T}} \frac{\mathbb{1}\left[\alpha^c_t = h^c_{i,t}\right]}{P_t(\alpha^c_t = h^c_{i,t})}(D_t(h^c_{i,t}) - h^e_{i,t}).$$

Thus, whenever $h^e_{i,T} > \tilde{\alpha}^e_T$, then by construction $w_t(i) < 0$, and therefore

$$\sum_{t \in B_{i,\leq T}} \frac{\mathbb{1}\left[\alpha^c_t = h^c_{i,t}\right]}{P(\alpha^c_t = h^c_{i,t})}(D_t(h^c_{i,t}) - h^e_{i,t}) \leq -\mathrm{CA}(T).$$

Thus, we get that among $T \in \tilde{B}_i$ (the times where t strictly outpromises the new estimates), the empirical record on the test set goes to $-\infty$.

It is left to show that overestimation remains low if we increase the estimates from α^e to $\tilde{\alpha}^e$. We have

$$\sum_{t=1}^{T} \tilde{\alpha}^e_t - D_t(h^c_{i,t}) = \sum_{t=1}^{T} \alpha^e_t - D_t(h^c_{i,t}) + \sum_{t=1}^{T} \tilde{\alpha}^e_t - \alpha^e_t.$$

The first sum is sublinear by assumption. So we only have to show that $\sum_{t=1}^{T} \tilde{\alpha}^e_t - \alpha^e_t$ is sublinear in T. We have

$$\sum_{t=1}^{T} \tilde{\alpha}^e_t - \alpha^e_t = \sum_i \sum_{t \in M_{i,\leq T}} h^e_{i,t} - \alpha^e_t. \tag{9}$$

So, it is left to show that the increase on behalf of each expert i is sublinear.

Now, we use IWE again. That is, we consider

$$\sum_{t \in M_{i,\leq T}} \frac{\mathbb{1}\left[\alpha^c_t = h^c_{i,t}\right]}{P(\alpha^c_t = h^c_{i,t})}(h^e_{i,t} - \alpha^e_t).$$

By the same argument as above, we can show that the difference between this term and $\sum_{t \in M_{i,\leq T}} h^e_{i,t} - \alpha^e_t$ is sublinear. So it is enough to show that this term is sublinear.

Now notice that

$$\begin{aligned} w_T(i) &= \mathrm{CA}(T) + \sum_{t \in M_{i,\leq T}} \frac{\mathbb{1}\left[\alpha^c_t = h^c_{i,t}\right]}{P(\alpha^c_t = h^c_{i,t})} \underbrace{(D_t(h^c_{i,t}) - h^e_{i,t})}_{=(D_t(h^c_{i,t}) - \alpha^e_t) - (h^e_{i,t} - \alpha^e_t)} + \sum_{t \in B_{i,\leq T} - M_i} \frac{\mathbb{1}\left[\alpha^c_t = h^c_{i,t}\right]}{\alpha^c_t = h^c_{i,t}} \left(D_t(h^c_{i,t}) - \alpha^e_t\right) \\ &= \mathrm{CA}(T) - \sum_{t \in M_{i,\leq T}} \frac{\mathbb{1}\left[\alpha^c_t = h^c_{i,t}\right]}{P_t(\alpha^c_t = h^c_{i,t})}(h^e_{i,t} - \alpha^e_t) + \sum_{t \in B_{i,\leq T}} \frac{\mathbb{1}\left[\alpha^c_t = h^c_{i,t}\right]}{P_t(\alpha^c_t = h^c_{i,t})}(D_t(h^c_{i,t}) - \alpha^e_t). \end{aligned}$$

Now, for $T \in M_i$, it must be $w_T(i) \geq 0$. Still, $w_T(i)$ can fall under 0, but only by $\hat{R}^m_t$ for some $t \in \{1, ..., T\}$, which is in $o(T)$. Thus,

$$\sum_{t \in M_{i,\leq T}} \frac{\mathbb{1}\left[\alpha^c_t = h^c_{i,t}\right]}{P_t(\alpha^c_t = h^c_{i,t})}(h^e_{i,t} - \alpha^e_t)\mathrm{CA}(T) + \sum_{t \in B_{i,\leq T}} \frac{\mathbb{1}\left[\alpha^c_t = h^c_{i,t}\right]}{P_t(\alpha^c_t = h^c_{i,t})}(D_t(h^c_{i,t}) - \alpha^e_t) + o(T)$$

CA is sublinear by construction and the second summand has been shown to be sublinear above. □

D. Why an even simpler theory fails and estimates are necessary

A simple mechanism of learning to choose is the *law of effect* (LoE) (Thorndike, 1911, p. 244):

> 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 (cf. Sutton & Barto, 1998, Section 1.6). 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, $\mathrm{DP}_t \in \mathrm{Fin}([0,1])$ with $D_t = \mathrm{id}$. 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 6.1). 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 $\bar{D}$ in which which this testing procedure introduces a relevant bias. In particular, the positive results we have proven in Theorems 6.1, 6.2 and 6.4 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 $\bar{D}$. 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 B.

E. Factoring team decisions

Theorem E.1. *Let $n \in \mathbb{N}$ be a positive natural number. Let $\bar{D}$ be a a decision problem sequence where every $t \in \mathbb{N}$, $\mathrm{DP}_t = \mathrm{DP}_{t,1} \times ... \times \mathrm{DP}_{t,n}$, for some sets $DP_{t,1}, ..., \mathrm{DP}_{t,n}$. Let $\bar{\alpha}$ be a BRIA for $\bar{D}$ covering the set of e.c. hypotheses. Now for any t let $((a_{t,1} \in DP_{t,1}, ..., a_{t,n} \in \mathrm{DP}_{t,n}), v_t) = \alpha_t$ in order to define $\alpha_{i,t} = (a_{t,i}, v_t)$ and*

$$D_{i,t} \colon \mathrm{DP}_{t,i} \to [0,1] : a'_{t,i} \mapsto D_t(a_{t,1}, ..., a_{t,i-1}, a'_{t,i}, a_{t,i+1}, ..., a_{t,n}).$$

for $i = 1, ..., n$. Then for $i = 1, .., n$, α_i is a BRIA for $\bar{D}_i$ covering the e.c. hypotheses.

Instead of considering sets DP_t that are already the Cartesian products of a bunch of sets, one could also factorize any given set (unless its number of elements is 1 or a prime number) (Garrabrant, 2021, Section 2). For example, a decision from $\{1, 2, 3, 4\}$ can be factorized into a decision of $\{1, 2\}$ versus $\{3, 4\}$, and a decision of $\{1, 3\}$ versus $\{2, 4\}$.

Proof. Low overestimation: Clearly,

$$\mathcal{L}(\bar{\alpha}_i, \bar{D}_i) = \sum_{t=1}^{T} \alpha_{i,t}^e - D_{i,t}(\alpha_{i,t}^c) = \sum_{t=1}^{T} \alpha_t^e - D_t(\alpha_t^c) \leq 0,$$

where the last step is by the assumption that $\bar{\alpha}$ is a BRIA and therefore does not overestimate in the limit.

Coverage: Let $\bar{h}_i$ be an e.c. hypothesis for $\bar{D}_i$. Let $\bar{h}$ be a hypothesis for $\bar{D}$ s.t. the i-th entry of h_t^c is equal to h_t^c, and $h_t^e = h_{i,t}^e$. Clearly, such an e.c. hypothesis exists. Let M be $\bar{\alpha}$'s test set for $\bar{h}$. We will also use M as $\bar{\alpha}_i$ s test set for $\bar{h}_i$. Also, let B be the set of times at which h_i outpromises $\bar{\alpha}_i$. Note that B is thereby also equal to the set of times at which $\bar{h}$ outbids $\bar{\alpha}$.

We now need to show that if B is infinite, then $(l_T(\bar{\alpha}_i, \bar{D}_i, M, \bar{h}_i))_{T \in B} \to -\infty$. To prove this, notice that for all T,

$$\begin{aligned} l_T(\bar{\alpha}_i, \bar{D}_i, M, \bar{h}_i) &= \sum_{t \in M_{\leq T}} D_{i,t}(h_{i,t}^c) - h_{i,t}^e \\ &= \sum_{t \in M_{\leq T}} D_{i,t}(\alpha_{i,t}^c) - h_{i,t}^e \\ &= \sum_{t \in M_{\leq T}} D_t(\alpha_t^c) - h_t^e \\ &= \sum_{t \in M_{\leq T}} D_t(h_t^c) - h_t^e \\ &= l_T(\bar{\alpha}, \bar{D}, M, \bar{h}). \end{aligned}$$

By assumption that $\bar{\alpha}$ is a BRIA, the final term goes to $-\infty$ within $T \in B$ if B is infinite. □

So if α is a BRIA, $\alpha_1, ..., \alpha_n$ are BRIAs. Note that the converse of this does not hold.

F. Dominance

Proposition F.1. *There is a decision process $\bar{D}$, a BRIA $\bar{\alpha}$ for the set of e.c. hypotheses and a positive number $\Delta > 0$ s.t. for all $t \in \mathbb{N}$, $a_t, b_t \in \mathrm{DP}_t$ $D_t(a_t) > D_t(b_t) + \Delta$ but with limit frequency 1 we have that $\alpha_t^c = b_t$.*

This is shown by Newcomb's problem (Appendix B). In fact, Newcomb's problem shows that for any algorithm that constructs BRIAs, there is a $\bar{D}$ s.t. the algorithm's BRIA converges to a_t.

Of course, various dominance-like results follow from the results of Section 6. However, more interesting applications of dominance are arguably ones where the conditions of these results aren't satisfied, e.g., where it is very unclear how one would assign expected utilities to different options. We will now give some reasons for why it's difficult to give any dominance result for BRIAs that does not follow from the results of Section 6.

The first thing to notice is that relationships such as $D_t(a_t) > D_t(b_t) + \Delta$ (for all t) are irrelevant for our theory, as shown by Newcomb's problem, SAO, etc. Instead, our dominance relation needs to be statistical and relative to the test set. Roughly, we must make an assumption that when testing a_t, the rewards are (on average) higher (by Δ) than the reward of taking b_t in rounds in which b_t is taken. Of course, this already means that the result will be quite different from traditional notions of dominance.

A second, subtler issue relates to the use of estimates in our theory.[6] To ensure that b_t is not taken with limit frequency,

[6] As noted in Appendix D, an alternative theory could simply require that an agent tests various choice policies and in the limit follows the ones that are empirically most successful. For such a theory, a condition like the one in the previous paragraph probably suffices.

we would need to ensure not only that the a_t-recommending hypothesis doesn't underperform on its test set (as described above). We also need to ensure that this hypothesis is tested on a set on which it doesn't overestimate. We therefore need a further assumption that gives us some way to safely and efficiently estimate a_t, e.g., based on past values of a_t, b_t or estimates α_t^e. While this assumption can be made in relatively sneaky ways, we have not found any particularly interesting version of this claim.

We now discuss a subtler issue that relates to the use of estimates in our theory to show why a particularly simple approach doesn't work. A first attempt might be to assume that for every test set M, $\mathrm{avg}_{t\in M_{\leq T}} D_t(a_t) > \Delta + \mathrm{avg}_{t\leq T\colon \alpha_t^c = b_t} \alpha_t^c$ as $T \to \infty$, where $\mathrm{avg}_{t\in N} f(t) := {}^1\!/\!_{|N|} \sum_{t\in N} f(t)$ for any finite set N and function f on N. That is, we assume that a_t performs better on any test set than b_t performs when taken by α. [TODO: finish this paragraph.] The trouble is that to obtain a conclusion we need to transform such an assumption into a hypothesis that not only recommends a_t (and thus receives relatively high rewards on average) but also makes appropriate estimates.

G. Schnorr bounded algorithmic randomness

Definition G.1. A *martingale* is a function $d\colon \mathbb{B}^* \to [0,\infty)$ s.t. for all $w \in$, $d(w) = {}^1\!/\!_2 d(w0) + {}^1\!/\!_2 d(w1)$. Let $w \in \mathbb{B}^\infty$ be an infinite sequence. We say that d succeeds on w if $\limsup_{n\to\infty} d(w_1 ... w_n) = \infty$.

Definition G.2. 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 G.3. *Let $\bar{D}$ be a decision problem sequence and α be an ($O(h(t))$-computable) BRIA for $\bar{D}$ covering all e.c. hypotheses. Let $\bar{a}$ be a sequence of terms in $\mathcal{T}$ s.t. $a_t \in \mathrm{DP}_t$ for all $t \in \mathbb{N}$ and the values $D_t(a_t) \in \{0,1\}$ are ($O(h(t))$-boundedly) Schnorr random. Then in the limit as $T \to \infty$, it holds that $\sum_{t=1}^T D_t(\alpha_t^c)/T \geq {}^1\!/\!_2$.*

Proof. We conduct a proof by contradiction. Assume that there is $\epsilon > 0$ s.t. $\sum_{t=1}^T D_t(\alpha_t^c)/T < {}^1\!/\!_2 - \epsilon$ infinitely often. Then by the no overestimation criterion, there must also be an $\epsilon > 0$ s.t. $\sum_{t=1}^T \alpha_t^e)/T < {}^1\!/\!_2 - \epsilon$. Consider the hypothesis $h_{a,\epsilon}$ that always estimates ${}^1\!/\!_2 - \epsilon$ and recommends a_t. Now let M_ϵ be $\bar{\alpha}$'s test for $h_{a,\epsilon}$. From the fact that $\bar{\alpha}$ rejects $h_{a,\epsilon}$ infinitely often, it follows that there are infinitely many $T \in \mathbb{N}$ such that $\sum_{t\in M_{\leq T}} D_t(a) - ({}^1\!/\!_2 - \epsilon) < 0$.

From this fact, we will now define an ($O(h(t))$-computable) martingale d that succeeds on the sequence $(D_t(a_t))_{t\in\mathbb{N}}$. To readers familiar with this literature, this will probably be familiar. First, define $d() = 0$. Whenever T is not in M, define $d((D_t(a_t))_{t<T}0) = d((D_t(a_t))_{t<T}) = d((D_t(a_t))_{t<T}1)$. That is, when $T \notin M$, don't bet on $D_T(a_T)$. If $T \in M$, then bet some small, constant fraction δ of the current money that the next bit will be 0. That is, $d((D_t(a_t))_{t<T}0) = (1+\delta)d((D_t(a_t))_{t<T})$ and $d((D_t(a_t))_{t<T}1) = (1-\delta)d((D_t(a_t))_{t<T})$. Clearly, d thus defined is a martingale.

Overall, we now know that there are infinitely many N s.t. for some T the wealth is $d((D_t(a_t))_{t\leq T}) \geq (1+\delta)^{N+\epsilon N}(1-\delta)^N$. It is left to show that for small enough δ, $(1+\delta)^{N+\epsilon N}(1-\delta)^N \to \infty$ as $N \to \infty$.

First notice that

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

So we need only show that for small enough but positive δ, $(1-\delta^2)(1+\delta)^\epsilon > 1$. 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 $\frac{d}{d\delta}(1-\delta^2)(1+\delta)^\epsilon = (1+\delta)^\epsilon(\epsilon - \delta(\epsilon+2))$. Inserting $\delta = 0$ yields ϵ, which is positive. □

H. A few minor results

In this section, we give a few minor results about the BRIA criterion. We don't use them anywhere, but they are helpful to understand what the BRIA criterion is about.

First, we simply note that the BRIA criterion becomes (weakly) stronger if we expand the set of hypotheses under consideration, which is immediate from the definitions in Section 4.

Proposition H.1. *Let $\mathbb{H}, \mathbb{H}'$ be sets of hypotheses such that $\mathbb{H}' \subseteq \mathbb{H}$. Then any BRIA for $\mathbb{H}$ is also a BRIA for $\mathbb{H}'$.*

The following result shows that if we change a BRIA's decisions and estimates for a finite number of decisions in $\bar{D}$, it remains a BRIA.

Proposition H.2. *Let $\bar{\alpha}$ be a BRIA for $\bar{D}$ covering $\mathbb{H}$. If for all but finitely many $t \in \mathbb{N}$ it is $\zeta_t = \alpha_t$, then $\bar{\zeta}$ is also BRIA for $\mathbb{H}$, $\bar{D}$.*

Proposition H.3. *Let $\bar{\alpha}$ be a BRIA covering h. Let h' be s.t. $h_t = h'_t$ for all but finitely many t. Then $\bar{\alpha}$ covers h'.*

Proposition H.4. *Let $\bar{\alpha}$ be a BRIA for $\bar{D}$ covering h. Let $f\colon \mathbb{N} \to \mathbb{N}$ be a bijection s.t. $f(n) - n$ is bounded (from above and below) (i.e., there is a number x s.t. $|f(n) - n| < x$ for all $n \in \mathbb{N}$). Then $\alpha_{f(1)}, \alpha_{f(2)}, ...$ is a BRIA for $D_{f(1)}, D_{f(2)}, ...$ covering $h_{f(1)}, h_{f(2)}, ...$*

Proposition H.5. *Let $\bar{\alpha}$ be a BRIA for $\bar{D}$ covering $\mathbb{H}$. Let $\bar{\epsilon}$ be a sequence of non-negative numbers such that $\sum_{t=1}^{T} \epsilon_t / T \to 0$ as $T \to \infty$. Let $\zeta_t = (\alpha_t^c, \alpha_t^e + \epsilon_t)$ for all t. Then $\bar{\zeta}$ is a BRIA for $\bar{D}$ covering $\mathbb{H}$.*

Note that decreasing estimates by a similar sequence $\bar{\epsilon}$ in general does not maintain the BRIA property. For example, if the estimates in rounds in which an option "0.5" is chosen is decreased below 0.5, the resulting agent would be exploitable by a hypothesis that recommends "0.5" and promises 0.5.