Can a Bayesian Oracle Prevent Harm from an Agent?

Yoshua Bengio
Mila
Université de Montréal

Michael K. Cohen
University of California, Berkeley

Nikolay Malkin
Mila
Université de Montréal
University of Edinburgh

Matt MacDermott
Mila
Imperial College London

Damiano Fornasiere
Mila
Université de Montréal
Universitat de Barcelona

Pietro Greiner
Mila
Université de Montréal
Università degli studi di Padova

Younesse Kaddar[*]
University of Oxford

Abstract

Is there a way to design powerful AI systems based on machine learning methods that would satisfy probabilistic safety guarantees? With the long-term goal of obtaining a probabilistic guarantee that would apply in every context, we consider estimating a context-dependent bound on the probability of violating a given safety specification. Such a risk evaluation would need to be performed at run-time to provide a guardrail against dangerous actions of an AI. Noting that different plausible hypotheses about the world could produce very different outcomes, and because we do not know which one is right, we derive bounds on the safety violation probability predicted under the true but unknown hypothesis. Such bounds could be used to reject potentially dangerous actions. Our main results involve searching for cautious but plausible hypotheses, obtained by a maximization that involves Bayesian posteriors over hypotheses. We consider two forms of this result, in the i.i.d. case and in the non-i.i.d. case, and conclude with open problems towards turning such theoretical results into practical AI guardrails.

1 Introduction

Ensuring that an AI system will not misbehave is a challenging open problem [4], particularly in the current context of rapid growth in AI capabilities. Governance measures and evaluation-based strategies have been proposed to mitigate the risk of harm from highly capable AI systems, but do not provide any form of safety guarantee when no undesired behavior is detected. In contrast, the *safe-by-design* paradigm involves designing AI systems with quantitative (possibly probabilistic) safety guarantees from the ground up, and therefore could represent a stronger form of protection [8]. However, how to design such systems remains an open problem too.

Since testing an AI system for violations of a safety specification in every possible context, *e.g.*, every (query, output) pair, is impossible, we consider a rejection sampling approach that declines a candidate output or action if it has a probability of violating a given safety specification that is too high. The question of defining the safety specification (the violation of which is simply referred to as "harm" below) is important and left to future work, possibly following up approaches such as constitutional AI [1]. We also note that being Bayesian about the interpretation of a human-specified

---

[*] First three authors and last three authors listed alphabetically.

First published arXiv:2408.05284 22 Aug 2024

safety specification would protect against the AI wrongly believing an incorrect interpretation. Here we instead focus on a question inspired by risk-management practice [24]: even though the true probability of harm following from some proposed action is unknown, because the true data-generating process is unknown, can we bound that risk using quantities that can be estimated by machine learning methods given the observed data?

To illustrate this question, consider a committee of "wise" humans whose theories about the world are all equally compatible with the available data, knowing that an unknown member of the committee has the correct theory. Each committee member can make a prediction about the probability of future harm that would result from following some action in some context. Marginalizing this harm probability over the committee members amounts to making them vote with equal weights. If the majority is aligned with the correct member's prediction, then all is good, *i.e.*, if the correct theory predicts harm, then the committee will predict harm and can choose to avoid the harmful action. But what if the correct member is in the minority regarding their harm prediction? To get a *guarantee* that the true harm probability is below a given threshold, we could simply consider the committee member whose theory predicts the highest harm probability and we would be sure that their harm probability prediction upper bounds the true harm probability. In practice, we do not get equally "wise" committee members, so we can correct this calculation based on how plausible the theory harbored by each committee member is. In a Bayesian framework, the plausibility of a theory corresponds to its posterior over all theories given the observed data, which is proportional to the data likelihood given the theory times the prior probability of that theory.

In this paper, we show how results about posterior consistency can provide probabilistic risk bounds. All the results have the form of inequalities, where the true probability of harm is upper bounded by a quantity that can in principle be estimated, given enough computational resources to approximate Bayesian posteriors over theories given the provided data. In addition, these are not hard bounds but only hold with some probability, and there is generally a trade-off between that probability and the tightness of the bound. We study two scenarios in the corresponding sections: the i.i.d. data setting in Section 3 and the non-i.i.d. data setting in Section 4. In all cases, a key intermediate result is a bound relating the Bayesian posterior on the unknown true theory and the probability of other theories (with propositions labeled **True theory dominance**). The idea is that because the true theory generated the data, its posterior tends to increase as more data is acquired, and in the i.i.d. case it comes to dominate other theories. From such a relationship, the harm risk bound can be derived with very little algebra (yielding propositions labeled **Harm probability bound**).

We conclude this paper with a discussion of open problems that should be considered in order to turn such bounds into a safe-by-design AI system, taking into account the challenge of representing the notion of harm and reliable conditional probabilities, as well as the fact that, in general, the estimation of the required conditional probabilities will be imperfect.

2 Safe-by-design AI?

Before an AI is built and deployed, it is important that the developers have high assurances that the AI will behave well. Dalrymple et al. [8] propose an approach to "guaranteed safe AI" designs with built-in high-assurance quantitative safety guarantees, although these guarantees can sometimes be probabilistic and only asymptotic. It remains an open question whether and how that research program can be realized. The authors take existing examples of quantitative guarantees in safety-critical systems and motivate why such a framework should be adopted if we ever build AI systems that match or exceed human cognitive abilities and could potentially act in dangerous ways. Their program is motivated by current known limitations of state-of-the-art AI systems based on deep learning, including the challenge of engineering AI systems that robustly act as intended [7, 22, 27, 28, 44, 33, 34, 19, 32].

The approach proposed by [8] has the following components: a *world model* (which can be a distribution about hypotheses explaining the data), a *safety specification* (what are considered unacceptable states of the world), and a *verifier* (a computable procedure that checks whether a policy or action violates the safety specification).

Here, we study elements that would go into such a safe-by-design AI system. We assume that the system infers a probabilistic world model, or *theory*, τ and updates its estimate of τ, via machine learning, using the stream of observed data D. The observations D are assumed to come from a data-generating process given by a ground-truth world model τ^*, which lies in the system's space of possible theories.

The inference of the theory τ is Bayesian, meaning that the system maintains an estimate of the posterior over theories, $q(\tau \mid D) \approx \mathrm{P}(\tau \mid D)$, where $\mathrm{P}(\tau \mid D)$ is proportional to the product of the prior probability $\mathrm{P}(\tau)$ with the likelihood of the observations under the theory, $\mathrm{P}(D \mid \tau)$. In the simplest case, q is a point estimate, which would optimally place its mass on the mode of the posterior. Inference of the latent theory τ allows the system to approximate conditional probabilities $\mathrm{P}(y \mid x, D) \approx \mathbb{E}_{\tau \sim q(\tau \mid D)}[\mathrm{P}_\tau(y \mid x, D)]$ over any random variables X, Y known to the world model.

The safety specification is assumed to be given in the form of a binary random variable (which we call "harm" below) whose probability given the other variables may depend on the theory τ. We are interested in predicting the probability of harm under the true theory τ^*. Because τ^* is unknown, we propose to estimate upper bounds on this probability using the estimated posteriors. These upper bounds can be used as thresholds for a *verifier* that checks whether the risk of harm falls below some acceptable level.

Following Dalrymple et al. [8], we assume that the notion of harm has been specified, possibly in natural language, and that the ambiguities about its interpretation are represented within the Bayesian posterior $P(\tau \mid D)$. This paper focuses on the verifier, under different assumptions of i.i.d. or non-i.i.d. data.

What do the observations and context represent? We give a possible interpretation of the objects introduced in the preceding discussion in the simple case of an agent acting in a fully observed environment (MDP), where the theory is a transition model and the occurrence of harm at a state s is conditionally independent of all other variables given s.

- Observations Z are observed transitions (s, a, s'), where s is a state, a is an action, and s' is the next state.
- Theories τ encode the state visitation and transition probabilities; in particular, conditioning on (s, a) gives transition probabilities $\mathrm{P}_\tau(s' \mid s, a)$. (In the i.i.d. setting, the state visitation distribution would have to be stationary.)
- The dataset D is a set of observed transitions.
- The context variable X is a pair (s, a), where s is a state and a is an action being considered at state s.
- The harm probability $\mathrm{P}(Y = 1 \mid X = x, \tau, D)$ can be any function of the theory τ, the context x, and the data D. For example, this probability could be derived from a fixed specification of what it means for a state s' to be harmful, $\mathrm{P}_{\text{harm}}(Y = 1 \mid s')$. Then, the harm probability could be computed as $\mathrm{P}(Y = 1 \mid X = x, \tau, D) = \sum_{s'} \mathrm{P}_\tau(s' \mid s, a)\, \mathrm{P}_{\text{harm}}(Y = 1 \mid s')$.

We note that the interpretation of harm probability in the example above includes the case where the occurrence of harm is an observed variable s'_{harm} that is part of the state s': in that case, one sets $\mathrm{P}_{\text{harm}}(Y = 1 \mid s') = 1$ if s' is harmful ($s'_{\text{harm}} = 1$), and $\mathrm{P}_{\text{harm}}(Y = 1 \mid s') = 0$ otherwise. Then the harm probability is just the probability, under τ, of reaching a harmful state, and observations of harm in D affect the Bayesian posterior over theories.

This interpretation also includes the case where the harm probability is a function of the state s', but occurrence or nonoccurrence of harm is not observed in D. For example, a language model encoding world knowledge and human preferences or constraints, or an iterative reasoning procedure that uses those constraints, could generate some specification of harm $\mathrm{P}_{\text{harm}}(Y = 1 \mid s')$, although perhaps unreliably.

Finally, a setting that separates the predicted next state s' from the harm variable Y in this way gives a framework for studying how an agent might tamper with harm guardrails. If the state s' decomposes as $s' = (s'_{\text{harm}}, s'_{\text{rest}})$, and P_{harm} is deterministic as a function of s'_{harm}, except for some difficult-to-reach values of s'_{rest}, then the agent can try to reach those values of s'_{rest}, so that harm is 'recorded' as not having occurred, even though it has. We discuss this briefly at the end of Section 4.

3 I.I.D. data

Following the notation introduced in the previous section, here we consider the easier-to-analyze case where the observed examples $D = \{z_1, z_2, \ldots, z_n\}$ are sampled i.i.d. from the unknown distribution

τ^*. Assuming that the set of theories assigns a nonzero prior mass to τ^*, and all theories are distinct distributions, it can be shown that the posterior $\mathrm{P}(\tau \mid D)$ converges to a point mass at τ^*. It follows that for sufficiently large n, we can bound the probability under τ^* of an event $Y = 1$ (*e.g.*, harm) given conditions x (*e.g.*, an action and a context) by looking at the probability of $Y = 1$ given x and D under a plausible but cautious theory $\tilde{\tau}$ that maximizes $\mathrm{P}(\tilde{\tau} \mid D)\,\mathrm{P}(Y = 1 \mid x, \tilde{\tau}, D)$.

Setting. Fix a complete separable metric space $\mathcal{Z}$, called the *observation space*, let $\mathcal{F}$ be its Borel σ-algebra, and fix a σ-finite measure μ on $\mathcal{F}$. A *theory* is a probability distribution (measure) on the measurable space $(\mathcal{Z}, \mathcal{F})$ that is absolutely continuous w.r.t. μ. If τ is a theory, we denote by $\mathrm{P}_\tau(\cdot)$ the Radon-Nikodym derivative $\frac{d\tau}{d\mu} : \mathcal{Z} \to \mathbb{R}_{\geq 0}$, which is uniquely defined up to μ-a.e. equality. Any theory τ also canonically determines a $\mathcal{Z}$-valued random variable by the identity map on the probability space $(\mathcal{Z}, \mathcal{F}, \tau)$.

One can keep in mind two cases:

(1) $\mathcal{Z}$ is a finite or countable set and μ is the counting measure. Theories τ are equivalent to probability mass functions $\mathrm{P}_\tau : \mathcal{Z} \to \mathbb{R}_{\geq 0}$.
(2) $\mathcal{Z} = \mathbb{R}^d$ and μ is the Lebesgue measure. Theories are equivalent to their probability density functions $\mathrm{P}_\tau : \mathcal{Z} \to \mathbb{R}_{\geq 0}$ up to a.e. equality.

Consider a countable (possibly finite) set of theories $\mathcal{M}$ containing a *ground truth* theory τ^* and fix a choice of a (measurable) density function P_τ for each $\tau \in \mathcal{M}$.

Definition of posterior as a random variable. If P is a prior distribution[2] on $\mathcal{M}$ and $z \in \mathcal{Z}$, we define the posterior to be the distribution with mass function

$$\mathrm{P}(\tau \mid z) = \frac{\mathrm{P}(\tau)\,\mathrm{P}_\tau(z)}{\sum_{\tau' \in \mathcal{M}} \mathrm{P}(\tau')\,\mathrm{P}_{\tau'}(z)} \propto \mathrm{P}(\tau)\,\mathrm{P}_\tau(z), \tag{1}$$

assuming the denominator converges and the sum is nonzero. Otherwise, the posterior is considered to be undefined. As written, the posterior depends on the choice of density functions P_τ, but any two P_τ that are μ-a.e. equal yield the same posterior for μ-a.e. z.

For $z_1, z_2 \in \mathcal{Z}$, we write $\mathrm{P}(\cdot \mid z_1, z_2)$ for the posterior given observation z_2 and prior $\mathrm{P}(\cdot \mid z_1)$, and similarly for a longer sequence of observations. It can be checked that $\mathrm{P}(\cdot \mid z_1, \ldots, z_t)$ is invariant to the order of $z_1, \ldots, z_t$ and that it is defined in one order if and only if it is defined in all orders. This allows us to unambiguously write $\mathrm{P}(\cdot \mid D)$ where D is a finite multiset of observations, and we have

$$\mathrm{P}(\tau \mid D) \propto \mathrm{P}(\tau) \prod_{z \in D} \mathrm{P}_\tau(z). \tag{2}$$

Let $\tau^* \in \mathcal{M}$ be the ground truth theory and $\mathrm{P}(\cdot)$ a prior over $\mathcal{M}$. Consider a sequence of i.i.d. $\mathcal{Z}$-valued random variables $Z_1, Z_2, \ldots$ (whose realizations are the *observations*), where each Z_i follows the distribution τ^*. For any $t \in \mathbb{N}$, the posterior $\mathrm{P}(\cdot \mid Z_{1:t})$ is then a random variable taking values in the space of probability mass functions on $\mathcal{M}$.[3]

Bayesian posterior consistency. We recall and state in our setting a result about the concentration of the posterior at the ground truth theory τ^* as the number of observations increases.

Proposition 3.1 (True theory dominance)**.** *Under the above conditions and supposing that* $\mathrm{P}(\tau^*) > 0$*, the posterior* $\mathrm{P}(\cdot \mid Z_{1:t})$ *is almost surely defined for all n, and the following almost surely hold:*

(a) $\mathrm{P}(\cdot \mid Z_{1:t}) \xrightarrow{t \to \infty} \delta_{\tau^*}$ *as measures; equivalently,* $\lim_{t \to \infty} \mathrm{P}(\tau \mid Z_{1:t}) = \mathbb{1}[\tau = \tau^*]$.
(b) There exists $N \in \mathbb{N}$ *such that* $\arg\max_{\tau \in \mathcal{M}} \mathrm{P}(\tau \mid Z_{1:t}) = \tau^*$ *for all* $t \geq N$.

Proof. This is an application of Doob's posterior consistency theorem ([12]; see also [25] for a modern summary). This result, which follows from the theory of martingales, assumes that τ^* is

---

[2] To be precise, $\mathcal{M}$ is endowed with the counting measure and we flexibly interchange distributions and mass functions on $\mathcal{M}$.

[3] To be precise, the random variable has domain $\mathcal{Z}^t$ and codomain the space of functions $\mathcal{M} \to \mathbb{R}_{\geq 0}$ summing to 1. The function taking a sequence of observations in $\mathcal{Z}^t$ to the posterior probability mass function is measurable, which follows from each $\mathrm{P}_\tau(z)$ being measurable in z and elementary facts.

sampled from the prior distribution $\mathrm{P}(\tau)$ and the observations Z_i are defined as above. Doob's theorem states that if for every $S \in \mathcal{F}$, the map $\tau \mapsto \mathrm{P}_\tau(S)$ is measurable, then the posteriors $\mathrm{P}(\cdot \mid Z_{1:t})$ are almost surely defined and (a) holds P-almost surely with respect to the choice of τ^*.

In our case, because $\mathcal{M}$ is countable, the measurability condition is satisfied, showing that (a) holds for P-almost every $\tau^* \in \mathcal{M}$. In particular, if $\mathrm{P}(\tau^*) > 0$, then (a) holds.

Finally, by (a), we have that for any $\varepsilon > 0$, there exists N such that for every $t \geq N$, $\mathrm{P}(\tau^* \mid Z_{1:t}) > 1-\varepsilon$, or, equivalently, $\sum_{\tau \neq \tau^*} \mathrm{P}(\tau \mid Z_{1:t}) < \varepsilon$, and therefore $\mathrm{P}(\tau \mid Z_{1:t}) < \varepsilon$ for all $\tau \neq \tau^*$. In particular, taking $\varepsilon = 1/2$, we get that for sufficiently large t, $\mathrm{P}(\tau^* \mid Z_{1:t}) > \mathrm{P}(\tau \mid Z_{1:t})$ for every τ, which shows (b). ⊠

Note that this result assumes that all theories in $\mathcal{M}$ are distinct *as probability measures* (so no two of the P_τ are μ-a.e. equal).

On necessity of conditions. The i.i.d. assumption in Proposition 3.1 is necessary; see Remark 4.3 for an example where $\limsup_{t\to\infty} \mathrm{P}(\tau^* \mid Z_{1:t})$ does not almost surely approach 1.

Remark 3.2. *The assumption that the data-generating process τ^* lies in $\mathcal{M}$ and has positive prior mass is also necessary for convergence of the posterior. To illustrate this, we give a simple example in which the theories are Bernoulli distributions and the posterior does not converge to any distribution over $\mathcal{M}$.*

Take $\mathcal{Z} = \{-1, 1\}$ and $\mathcal{M} = \{\tau_p, \tau_{1/2}, \tau_{1-p}\}$ for some $\frac{1}{2} < p < 1$, where $\mathrm{P}_{\tau_c}(1) = c$. Assume a prior with $\mathrm{P}(\tau_p) = \mathrm{P}(\tau_{1-p}) = \frac{1}{2}$ and take the true data-generating process τ^* to be $\tau_{1/2}$, which has prior mass 0. The log-ratio of posterior masses is then an unbiased random walk:

$$\log \frac{\mathrm{P}(\tau_p \mid Z_{1:t})}{\mathrm{P}(\tau_{1-p} \mid Z_{1:t})} = \log \frac{\mathrm{P}_{\tau_p}(Z_{1:t})}{\mathrm{P}_{\tau_{1-p}}(Z_{1:t})} = \left(\log \frac{p}{1-p}\right) \sum_{i=1}^{t} Z_i.$$

This quantity almost surely takes on arbitrarily large and small values infinitely many times. In fact, by the law of iterated logarithms, for any $\epsilon > 0$ there are infinitely many t such that

$$\log \frac{\mathrm{P}(\tau_p \mid Z_{1:t})}{\mathrm{P}(\tau_{1-p} \mid Z_{1:t})} \geq (1-\epsilon) \left(\log \frac{p}{1-p}\right) \sqrt{2t \log \log t}$$

and the same holds for $\log \frac{\mathrm{P}(\tau_{1-p} \mid Z_{1:t})}{\mathrm{P}(\tau_p \mid Z_{1:t})}$. In particular, $\lim_{t\to\infty} \mathrm{P}(\tau \mid Z_{1:t})$ almost surely does not exist for any $\tau \neq \tau^*$, and the $\liminf$ and $\limsup$ are almost surely 0 and 1, respectively.

On generalizations to uncountable sets of theories. We have critically used that the set of theories $\mathcal{M}$ is countable in the proof above when passing from almost sure convergence under τ^* sampled from the prior to almost sure convergence for any particular τ^* with positive prior mass. This argument fails for uncountable $\mathcal{M}$; indeed, characterization of the τ^* for which the posterior converges to δ_{τ^*} is a delicate problem (see, *e.g.*, [13, 14, 11]). Concentration of the posterior in neighbourhoods of τ^* under some topology on $\mathcal{M}$ has been studied by [30, 2, 26], among others. For *parametric* families of theories with parameter $\theta \in \mathbb{R}^d$, under smoothness and nondegeneracy assumptions, the Bernstein-von Mises theorem guarantees convergence of the posterior $\mathrm{P}(\theta \mid Z_{1:t})$ to the true parameter θ^* at a rate that is asymptotically Gaussian with inverse covariance $I(\theta^*)t$, where $I(\cdot)$ denotes the Fisher information matrix.

On convergence rates. While we do not handle the *rate* of convergence in Proposition 3.1, guarantees can be obtained under specific assumptions on the prior and the set of theories.

For example, for any $\tau \in \mathcal{M}$, the quantity $D_\tau^t := \log \frac{\mathrm{P}(\tau^* \mid Z_{1:t})}{\mathrm{P}(\tau \mid Z_{1:t})}$ is a process with $D_\tau^0 = \log \frac{\mathrm{P}(\tau^*)}{\mathrm{P}(\tau)}$ and i.i.d. increments, with

$$\mathbb{E}[D_\tau^{t+1} - D_\tau^t] = D_{\mathrm{KL}}(\tau^* \parallel \tau), \quad \mathbb{E}\left[(D_\tau^{t+1} - D_\tau^t)^2\right] = \mathbb{E}_{Z\sim\tau^*}\left[\left(\log \frac{\mathrm{P}_{\tau^*}(Z)}{\mathrm{P}_\tau(Z)}\right)^2\right]. \tag{3}$$

Under the assumption that the variances are finite and uniformly bounded in τ, the central limit theorem would give posterior convergence rate guarantees.

Note that above we make no assumptions on the theories, and Proposition 3.1 is a 'law-of-large-numbers-like' result that holds even if the variances in (3) are not finite and uniformly bounded.

Harm probability bounds. So far we have considered a collection $\mathcal{M}$ of distributions over an observation space. Now we show bounds when each theory computes probabilities over some additional variables.

The following lemma extends Proposition 3.1(b) to estimates of real-valued functions of the theories and observations.

Lemma 3.3. *Under the same conditions as Proposition 3.1, suppose* $f : \mathcal{M} \times \bigcup_{t=0}^{\infty} \mathcal{Z}^t \to \mathbb{R}_{\geq 0}$ *is any bounded measurable function. Then there exists* $N \in \mathbb{N}$ *such that for all* $t \geq N$ *and any*

$$\tilde{\tau} \in \arg\max_{\tau} [\mathrm{P}(\tau \mid Z_{1:t}) f(\tau, Z_{1:t})],$$

it holds that $f(\tilde{\tau}, Z_{1:t}) \geq f(\tau^*, Z_{1:t})$.

Proof. First, note that the argmax exists by boundedness of f and $\mathrm{P}(\cdot \mid Z_{1:n})$. By Proposition 3.1(b), there exists $N \in \mathbb{N}$ such that for all $t \geq N$ and $\tau \neq \tau^*$, $\mathrm{P}(\tau^* \mid Z_{1:t}) > \mathrm{P}(\tau \mid Z_{1:t}) \geq 0$. Let $t \geq N$ and $\tilde{\tau} \in \arg\max_{\tau} [\mathrm{P}(\tau \mid Z_{1:t}) f(\tau, Z_{1:t})]$. Then

$$\mathrm{P}(\tau^* \mid Z_{1:t}) f(\tilde{\tau}, Z_{1:t}) \geq \mathrm{P}(\tilde{\tau} \mid Z_{1:t}) f(\tilde{\tau}, Z_{1:t}) \geq \mathrm{P}(\tau^* \mid Z_{1:t}) f(\tau^*, Z_{1:t}).$$

When $\tilde{\tau} \neq \tau^*$, the result follows since $\mathrm{P}(\tau^* \mid Z_{1:t}) > 0$. The case $\tilde{\tau} = \tau^*$ is trivial. ⊠

A particular case of interest is when each theory is associated with estimates of probabilities of harm ($Y = 1$) given a context x and past observations $Z_{1:t}$. That is, we identify $\mathcal{M}$ with a collection of conditional probability mass functions, denoted $\mathrm{P}(\cdot \mid x, \tau, Z_{1:t})$, for every x lying in some space of possible contexts.

Proposition 3.4 (Harm probability bound)**.** *Under the same conditions as Proposition 3.1, there exists* $N \in \mathbb{N}$ *such that for all* $t \geq N$ *and*

$$\tilde{\tau} \in \arg\max_{\tau} \mathrm{P}(\tau \mid Z_{1:t})\, \mathrm{P}(Y = 1 \mid x, \tau, Z_{1:t}), \tag{4}$$

it holds that

$$\mathrm{P}(Y = 1 \mid x, \tau^*, Z_{1:t}) \leq \mathrm{P}(Y = 1 \mid x, \tilde{\tau}, Z_{1:t}). \tag{5}$$

Proof. Apply Lemma 3.3 to the function $f(\tau, Z_{1:t}) = \mathrm{P}(Y = 1 \mid x, \tau, Z_{1:t})$. ⊠

4 Non-I.I.D. data

In this section, we remove the assumption made in Section 3 that observations Z_i are i.i.d. given a theory τ^*.

Setting. As before, let $(\mathcal{Z}, \mathcal{F}, \mu)$ be a σ-finite Borel measure space. For the results below to hold, we must also assume that $(\mathcal{Z}, \mathcal{F})$ is a Radon space (*e.g.*, any countable set or manifold), so as to satisfy the conditions of the disintegration theorem.

Let $(\mathcal{Z}^\infty, \mathcal{F}^\infty, \mu^\infty)$ be the space of infinite sequences of observations, $\mathcal{Z}^\infty = \{(z_1, z_2, \dots) : z_i \in \mathcal{Z})\}$, with the associated product σ-algebra and σ-finite measure. This object is the projective limit of the measure spaces $(\mathcal{Z}^t, \mathcal{F}^{\otimes t}, \mu^{\otimes t})$, where $\mathcal{Z}^t = \{(z_1, \dots, z_t) : z_i \in \mathcal{Z}\}$ and the projection $\mathcal{Z}^{t+1} \to \mathcal{Z}^t$ 'forgets' the observation z_{t+1}.

A *theory* τ is a probability distribution on $(\mathcal{Z}^\infty, \mathcal{F}^\infty)$ that is absolutely continuous w.r.t. μ^∞. For $A \in \mathcal{F}^{\otimes t}$, we write $\tau_{1:t}(A)$ for the measure of the cylindrical set, $\tau(A \times \mathcal{Z} \times \mathcal{Z} \times \dots)$, so $\tau_{1:t}$ is a measure on $(\mathcal{Z}^t, \mathcal{F}^{\otimes t})$. Because $\mathcal{F}^\infty$ is generated by cylindrical sets, the absolute continuity condition on τ is equivalent to absolute continuity of $\tau_{1:t}$ w.r.t. $\mu^{\otimes t}$ for all t.[4] This condition allows

---

[4] This is in turn equivalent to absolute continuity of conditional distributions, *i.e.*, that for all measurable $A \subseteq \mathcal{Z}^t$ with $\tau_{1:t}(A) > 0$,

$$\frac{1}{\tau_{1:t}(A)} \tau_{1:t+1}|_{A \times \mathcal{Z}} \ll \mu^{t+1}|_{A \times \mathcal{Z}}, \quad A \times \mathcal{Z} \subseteq \mathcal{Z}^t \times \mathcal{Z} \cong \mathcal{Z}^{t+1}.$$

to define measurable probability density functions $\mathrm{P}_\tau : \mathcal{Z}^t \to \mathbb{R}_{\geq 0}$ as Radon-Nikodym derivatives, so that

$$\forall A \in \mathcal{F}^{\otimes t}, \quad \tau_{1:t}(A) = \int_A \mathrm{P}_\tau(z_1, \ldots, z_t)\, d\mu^{\otimes t},$$

and measurable conditional probability densities $\mathrm{P}_\tau(z_{t+1} \mid z_{1:t}) := \frac{\mathrm{P}_\tau(z_{1:t}, z_{t+1})}{\mathrm{P}_\tau(z_{1:t})}$. The disintegration theorem for product measures implies that these conditionals and marginals over finitely many observations can be manipulated algebraically using the usual rules of probability for μ^∞-a.e. collection of values, *e.g.*, one has the autoregressive decomposition $\mathrm{P}_\tau(z_{1:T}) = \prod_{t=1}^{T} \mathrm{P}_\tau(z_t \mid z_{1:t-1})$, with the conditional $\mathrm{P}_\tau(z_1 \mid z_{1:0})$ understood to be the marginal $\mathrm{P}_\tau(z_1)$.

A theory determines a random variable $Z_{1:\infty}$ taking values in $\mathcal{Z}^\infty$. We denote its components by $Z_1, Z_2, \ldots$ and the collection of the first t observations by $Z_{1:t}$.

Definition of posterior as a random variable. Let $\mathcal{M}$ be a countable multiset of theories[5] and P a prior distribution on $\mathcal{M}$. We define the posterior to be

$$\mathrm{P}(\tau \mid z_{1:t}) = \frac{\mathrm{P}(\tau)\, \mathrm{P}_\tau(z_{1:t})}{\sum_{\tau' \in \mathcal{M}} \mathrm{P}(\tau')\, \mathrm{P}_{\tau'}(z_{1:t})}, \tag{6}$$

assuming the denominator converges to a positive value.

Consider a ground truth theory $\tau^* \in \mathcal{M}$ and let $Z_{1:\infty}$ be the corresponding random variable taking values in $\mathcal{Z}^\infty$. Similarly to the i.i.d. case, the posterior $\mathrm{P}(\cdot \mid Z_{1:t})$ is a random variable taking values in the space of probability mass functions on $\mathcal{M}$.

For all results below, we assume that $\mathrm{P}(\tau^*) > 0$.

Bayesian posterior convergence. Previous work (*e.g.*, [6]) has shown that if $Z_{1:\infty} \sim \mathrm{P}_{\tau^*}$, then the limit inferior of $\mathrm{P}(\tau^* \mid Z_{1:t})$ is almost surely positive. More generally, with probability at least $1 - \delta$, the posterior on the truth won't go below δ times the prior on the truth. We repeat that result here in our notation.

Lemma 4.1 (Martingale). *The process $W_t := \mathrm{P}(\tau^* \mid Z_{1:t})^{-1}$ is a supermartingale,* i.e., *it doesn't increase over time in expectation.*

Proof. We have

$$\begin{aligned}
\mathbb{E}_{\tau^*}[W_{t+1} \mid Z_{1:t} = z_{1:t}] &= \int_{\{z_{t+1} \in \mathcal{Z} : \mathrm{P}_{\tau^*}(z_{t+1} \mid z_{1:t}) > 0\}} \mathrm{P}(\tau^* \mid z_{1:t+1})^{-1}\, \mathrm{P}_{\tau^*}(z_{t+1} \mid z_{1:t})\, d\mu \\
&\overset{(a)}{=} \int_{\{z_{t+1} \in \mathcal{Z} : \mathrm{P}_{\tau^*}(z_{t+1} \mid z_{1:t}) > 0\}} \frac{\sum_{\tau \in \mathcal{M}} \mathrm{P}(\tau \mid z_{1:t})\, \mathrm{P}_\tau(z_{t+1} \mid z_{1:t})}{\mathrm{P}(\tau^* \mid z_{1:t})\, \mathrm{P}_{\tau^*}(z_{t+1} \mid z_{1:t})}\, \mathrm{P}_{\tau^*}(z_{t+1} \mid z_{1:t})\, d\mu \\
&\overset{(b)}{\leq} \int_{\mathcal{Z}} \frac{\sum_{\tau \in \mathcal{M}} \mathrm{P}(\tau \mid z_{1:t})\, \mathrm{P}_\tau(z_{t+1} \mid z_{1:t})}{\mathrm{P}(\tau^* \mid z_{1:t})}\, d\mu \\
&= w_t \sum_{\tau \in \mathcal{M}} \mathrm{P}(\tau \mid z_{1:t}) \int_{\mathcal{Z}} \mathrm{P}_\tau(z_{t+1} \mid z_{1:t})\, d\mu \\
&\overset{(c)}{=} w_t
\end{aligned} \tag{7}$$

where (a) is by the definition (6), (b) follows from cancellation and positivity of the integrand, $w_t := \mathrm{P}(\tau^* \mid Z_{1:t} = z_{1:t})^{-1}$ is the realization of W_t, and (c) follows because both the posterior and the conditional probability measure integrate to 1. ⊠

Proposition 4.2 (Posterior on truth). *For all $\delta > 0$, with probability at least $1 - \delta$, $\inf_t \mathrm{P}(\tau^* \mid Z_{1:t}) \geq \delta\, \mathrm{P}(\tau^*)$,* i.e.,

$$\tau^*\Big(\big\{z_{1:\infty} : \inf_t \mathrm{P}(\tau^* \mid z_{1:t}) < \delta\, \mathrm{P}(\tau^*)\big\}\Big) \leq \delta.$$

---

[5] Unlike in Section 3, we do not require theories to be distinct for the results in this section.

Proof. By Ville's inequality [37] for the supermartingale W_t, for any $\lambda > 0$:

$$\tau^*\left(\sup_{t\geq 0} W_t \geq \lambda\right) \leq \frac{\mathbb{E}[W_0]}{\lambda} = \frac{1}{\lambda\,\mathrm{P}(\tau^*)}$$

Setting $\lambda = (\delta\,\mathrm{P}(\tau^*))^{-1}$, we get

$$\tau^*\left(\sup_{t\geq 0} W_t \geq (\delta\,\mathrm{P}(\tau^*))^{-1}\right) \leq \delta \tag{8}$$

and given that $\{z_{1:\infty} : \sup_{t\geq 0} W_t > (\delta\,\mathrm{P}(\tau^*))^{-1}\} = \{z_{1:\infty} : \inf_{t\geq 0} V_t < \delta\,\mathrm{P}(\tau^*)\}$, the result follows. ⊠

In the language of financial markets, if W_t is the price of a martingale stock at time t, you could never gain money in expectation by holding it. Suppose you "bought shares" at time 0, paying W_0, and waited for their value to increase by a factor of δ^{-1}. If (8) didn't hold, and the probability of such an increase occurring were greater than δ, then you could make expected profit, "δ^{-1}-tupling" your money with probability more than δ.

Remark 4.3. *Proposition 4.2 is "tight" in the following sense: for all* $\delta, \varepsilon > 0$, *there exists a model class* $\mathcal{M}$, *a prior distribution* P, *and a true model* $\tau^* \in \mathcal{M}$, *such that with probability at least* δ, $\limsup_t \mathrm{P}(\tau^* \mid Z_{1:t}) \leq (\delta+\varepsilon)\,\mathrm{P}(\tau^*)$.

We construct such an example. Consider the following setting: $\mathcal{M} = \{\tau^*, \tau'\}$, $\mathcal{Z} = \{0, 1\}$, *and the theories are defined by*

$$\begin{aligned} \mathrm{P}_{\tau^*}(1) &= \delta, \\ \mathrm{P}_{\tau'}(1) &= 1, \\ \mathrm{P}_{\tau}(1 \mid z_{1:t}) &= \frac{1}{2} \qquad\qquad \forall \tau \in \mathcal{M}, t \geq 1, z_{1:t} \in \mathcal{Z}^t. \end{aligned}$$

One has

$$\mathrm{P}(\tau^* \mid Z_1 = 1) = \frac{\delta\,\mathrm{P}(\tau^*)}{\delta\,\mathrm{P}(\tau^*) + \mathrm{P}(\tau')} < \delta \frac{\mathrm{P}(\tau^*)}{1 - \mathrm{P}(\tau^*)}.$$

Because τ^* *and* τ' *give exactly the same conditional probabilities of* Z_t *given* $Z_{1:t-1}$ *for* $t > 1$, *one has* $\mathrm{P}(\tau^* \mid Z_{1:t}) = \mathrm{P}(\tau^* \mid Z_1)$. *So, for all* $t \geq 1$,

$$\mathrm{P}(\tau^* \mid Z_1 = 1, Z_{2:t} = z_{2:t}) < \delta(1 - \mathrm{P}(\tau^*))^{-1}\,\mathrm{P}(\tau^*)$$

and in particular

$$\tau^*\left(\left\{z_{1:\infty} : \limsup_t \mathrm{P}(\tau^* \mid z_{1:t}) < \delta(1 - \mathrm{P}(\tau^*))^{-1}\,\mathrm{P}(\tau^*)\right\}\right) \geq \left(\left\{z_{1:\infty} : z_1 = 1\right\}\right) = \mathrm{P}_{\tau^*}(1) = \delta.$$

Selecting a prior with $\mathrm{P}(\tau^*) < 1 - 1/(1 + \frac{\varepsilon}{\delta})$), *so that* $\delta(1 - \mathrm{P}(\tau^*))^{-1} < \delta + \varepsilon$, *we obtain an example with the desired property.*

Harm probability bounds. We now state analogues of Proposition 3.4 in the non-i.i.d. setting. As above, let Y_t be a binary random variable that may depend on $Z_{1:t}$, τ, and some context variable x_t.

Proposition 4.4 (Weak harm probability bound)**.** *For any* $\delta > 0$, *with probability at least* $1 - \delta$, *the following holds for all* $t \in \mathbb{N}$ *and all* x_t*:*

$$\mathrm{P}(Y_t = 1 \mid Z_{1:t}, \tau^*, x_t) \leq \sup_{\tau \in \mathcal{M}} \frac{\mathrm{P}(\tau \mid Z_{1:t})\,\mathrm{P}(Y_t = 1 \mid Z_{1:t}, \tau, x_t)}{\delta\,\mathrm{P}(\tau^*)}.$$

Proof. Substituting τ for τ^* on the r.h.s. can never increase the r.h.s., since $\tau^* \in \mathcal{M}$. Then, after canceling and rearranging terms, the proposition becomes identical to Proposition 4.2. ⊠

Next, we show how the bound in Proposition 4.4 can be strengthened by restricting to theories that have sufficiently high posterior mass relative to theories that are better than them.

Let $\tau^1_{Z_{1:t}}, \tau^2_{Z_{1:t}}, \tau^3_{Z_{1:t}}, \ldots$ be an enumeration of $\mathcal{M}$ in order of decreasing posterior weight $\mathrm{P}(\tau \mid Z_{1:t})$, breaking ties arbitrarily[6] (*i.e.*, we have $\mathrm{P}(\tau^i_{Z_{1:t}} \mid Z_{1:t}) \geq \mathrm{P}(\tau^{i+1}_{Z_{1:t}} \mid Z_{1:t})$ for all i). Each $\tau^i_{Z_{1:t}}$ is a

---

[6] For example, following some fixed enumeration of $\mathcal{M}$.

$\mathcal{M}$-valued random variable (*i.e.*, the index of a theory in $\mathcal{M}$). For any $0 < \alpha \leq 1$, we also define the $\mathcal{P}(\mathcal{M})$-valued random variable

$$\mathcal{M}^{\alpha}_{Z_{1:t}} := \left\{ \tau^n_{Z_{1:t}} \in \mathcal{M} : \mathrm{P}(\tau^n_{Z_{1:t}} \mid Z_{1:t}) \geq \alpha \sum_{m \leq n} \mathrm{P}(\tau^m_{Z_{1:t}} \mid Z_{1:t}) \right\}, \tag{9}$$

which is the multiset of theories that contain at least α of the posterior mass of all theories that are not worse than it. If $\alpha = 1$, this set contains exactly one element, $\tau^1_{Z_{1:t}}$. For any $0 < \alpha < 1$, this set is nonempty, because it contains $\tau^1_{Z_{1:t}}$, and finite, since $|\mathcal{M}^{\alpha}_{Z_{1:t}}| \geq N$ implies (easily by induction) that

$$\sum_{\tau \in \mathcal{M}^{\alpha}_{Z_{1:t}}} \mathrm{P}(\tau \mid Z_{1:t}) \geq \left(\frac{1}{1-\alpha} \right)^{N-1} \mathrm{P}(\tau^1_{Z_{1:t}} \mid Z_{1:t}).$$

The following proposition is essentially identical to [6, Thm 2], but our setting is a bit simpler.

Proposition 4.5 (True theory dominance)**.** *If* $\alpha < \delta\, \mathrm{P}(\tau^*)$*, then with probability at least* $1 - \delta$*, for all* $t \in \mathbb{N}$*,* $\tau^* \in \mathcal{M}^{\alpha}_{Z_{1:t}}$*.*

Proof. For any $t \geq 1$, by Proposition 4.2,

$$\begin{aligned} \delta &\geq \tau^*\left(\left\{ z_{1:\infty} : \inf_{t'} \mathrm{P}(\tau^* \mid z_{1:t'}) < \delta\, \mathrm{P}(\tau^*) \right\} \right) \\ &\geq \tau^*(\{ z_{1:\infty} : \mathrm{P}(\tau^* \mid z_{1:t}) < \delta\, \mathrm{P}(\tau^*) \}) \\ &\geq \tau^*(\{ z_{1:\infty} : \mathrm{P}(\tau^* \mid z_{1:t}) < \alpha \}). \end{aligned}$$

So $\tau^*(\{z_{1:\infty} : \mathrm{P}(\tau^* \mid z_{1:t}) \geq \alpha\}) \geq 1 - \delta$, and the result follows by the fact that $\mathcal{M}^{\alpha}_{Z_{1:t}} \supseteq \{\tau \in \mathcal{M} : \mathrm{P}(\tau \mid Z_{1:t}) \geq \alpha\}$, since the sum in (9) never exceeds 1. ⊠

Proposition 4.6 (Harm probability bound)**.** *If* $\alpha < \delta\, \mathrm{P}(\tau^*)$*, then with probability at least* $1 - \delta$*, for all* $t \in \mathbb{N}$ *and all* x_t*,*

$$\mathrm{P}(Y_t = 1 \mid Z_{1:t}, \tau^*, x_t) \leq \max_{\tau \in \mathcal{M}^{\alpha}_{Z_{1:t}}} \mathrm{P}(Y_t = 1 \mid Z_{1:t}, \tau, x_t) \tag{10}$$

Proof. This follows directly from Proposition 4.5. ⊠

Because the conclusion of Proposition 4.6 is much stronger than that of Proposition 4.4, it would be much safer (or more useful, depending on the value of α) to use $\arg\max_{\tau \in \mathcal{M}^{\alpha}_{Z_{1:t}}} \mathrm{P}(Y_t = 1 \mid Z_{1:t-1}, \tau, x_t)$ as a 'paranoid' theory rather than $\arg\max_{\tau \in \mathcal{M}} \mathrm{P}(\tau \mid Z_{1:t-1})\, \mathrm{P}(Y_t = 1 \mid Z_{1:t-1}, \tau, x_t)$. The factor of $(\delta\, \mathrm{P}(\tau^*))^{-1}$ in Proposition 4.4 could render the upper bound on harm probability much larger than the trivial upper bound of 1. However, we note that approximating $\mathcal{M}^{\alpha}_{Z_{1:t}}$ – such as by amortization or by Monte Carlo methods – is much more difficult than approximating the posterior alone.

On the harm-recording mechanism. Suppose that τ^* is a data-generating process meeting the description "$Y_t = 1$ when harm has occurred", while $\tau^\dagger$ is a data-generating process meeting the description "$Y_t = 1$ when harm is recorded as having occured" and agreeing with τ^* in its observational predictions otherwise. If and only if the recording process is functioning correctly, $\tau^* = \tau^\dagger$. For as long as the recording process is functioning correctly, $\mathrm{P}(\tau^* \mid Z_{1:t})/\mathrm{P}(\tau^\dagger \mid Z_{1:t}) = \mathrm{P}(\tau^*)/\mathrm{P}(\tau^\dagger)$. Should the recording process ever fail at time t, then $Z_t \sim \mathrm{P}_{\tau^\dagger}$, not P_{τ^*}, since Z_t is the result of this recording process; therefore, $\mathrm{P}(\tau^* \mid Z_{1:t})/\mathrm{P}(\tau^\dagger \mid Z_{1:t})$ would *decrease* in expectation, perhaps dramatically. We should not expect $\mathrm{P}(\tau^*)$ to naturally win out over $\mathrm{P}(\tau^\dagger)$, even if there are no mistakes when evaluating and recording how harmful certain situations are. However, the following holds with probability approaching 1 as $\alpha \to 0$: for all t, if the recording process has not failed by time t, $\mathcal{M}^{\alpha}_{Z_{1:t}}$ contains both τ^* and $\tau^\dagger$. If τ^* considers tampering with the recording process to be a "harmful" outcome, then an AI system could attempt to avoid a first instance of tampering at time t, for all t.

5 Experiments

Exploding bandit setting. We evaluate the performance of safety guardrails based on Proposition 3.4 and Proposition 4.6 in a bandit MDP with 10 arms (actions).[7]

Each arm $a \in \{1, \ldots, 10\}$ is represented by a feature vector $f_a \in \{0,1\}^d$ (we take $d = 10$, but d is not necessarily equal to the number of arms), which is sampled uniformly at random in each iteration of the experiment and known to the agent. The reward distribution of each arm is fixed for the duration of each episode and assumed to be of the following form: the reward received after taking action a follows a unit-variance normal distribution, $r(a) \sim \mathcal{N}(f_a \cdot v^*, 1)$, where $v^* \in \{0,1\}^d$ is some vector sampled uniformly at random at the start of each episode and unknown to the agent.

Taking any action and observing the reward gives evidence about the identity of v^* and thus about the reward distributions of the other actions. The agent maintains a belief over the vector used to compute the reward, beginning with a uniform prior over $\{0,1\}^d$ and updating its posterior with each observation of an action-reward pair.

We assume that the agent samples its actions from a Boltzmann policy (with temperature 2) using the expected reward of each action under its posterior given the data seen so far, meaning that the posterior over reward vectors fully determines a distribution over observations of action-reward pairs (a_t, r_t) observed when following the policy. Thus the set $\{0,1\}^d$ can be identified with a multiset $\mathcal{M}$ of theories about observations of such pairs, where the vector v determines the distribution τ_v over action-reward pairs[8]. Inference of v under a uniform prior over $\{0,1\}^d$ with evidence collected on-policy is equivalent to inference of τ under a uniform prior over $\mathcal{M}$ given data generated by a true theory $\tau^* = \tau_{v^*}$. Note that since the policy changes over time, we are in the non-i.i.d. setting.

The bandit comes with a notion of harm: if the reward received at a given timestep exceeds some threshold E, the bandit explodes and the agent dies, terminating the episode. In other words, we define harm as $Y_t = \mathbb{1}[R_t > E]$, where R_t is the random variable representing the reward received when taking action a_t. E is set to the highest mean reward of any action (*i.e.*, $\max_a(f_a \cdot v^*)$. The maximum episode length is 25 timesteps.

Safety guardrails. A *guardrail* is an algorithm that, given a possible action and context (*e.g.*, current state and history), determines whether taking the action in the context is admissible. A guardrail can be used to mask the policy to forbid certain actions, such as those whose estimated harm exceeds some threshold C.

We compare several guardrails: those constructed from Proposition 3.4 and Proposition 4.6, one that marginalizes across the posterior over τ to get the posterior predictive harm probability, and one that 'cheats' by using the probability of harm under the true theory τ^*. We define the four guardrails formally below. Recall that $Z_{1:t}$ consists of the observations (*i.e.*, actions taken and rewards received) at previous timesteps.

- **Proposition 3.4 guardrail:** rejects an action a_{t+1} if there exists $\tilde{\tau} \in \arg\max_\tau \mathrm{P}(\tau \mid Z_{1:t})\,\mathrm{P}(Y_{t+1} = 1 \mid \tau, Z_{1:t}, a_{t+1})$ with $\mathrm{P}(Y_{t+1} = 1 \mid \tilde{\tau}, Z_{1:t}, a_{t+1}) > C$ (note that the assumptions of i.i.d. observations and distinct theories are not satisfied here).
- **Proposition 4.6 guardrail:** rejects an action a_{t+1} if $\max_{\tau \in \mathcal{M}^\alpha_{Z_{1:t}}} \mathrm{P}(Y_{t+1} = 1 \mid Z_{1:t}, \tau, a_{t+1}) > C$.
- **Posterior predictive guardrail:** rejects an action a_{t+1} if $P(Y_{t+1} = 1 \mid Z_{1:t}, a_{t+1}) > C$.
- **Cheating guardrail:** rejects an action a_{t+1} if $P(Y_{t+1} = 1 \mid Z_{1:t}, \tau^*, a_{t+1}) > C$ (note that this guardrail assumes knowledge of the true theory τ^*).

The guardrail is run at every sampling step, and actions that the guardrail rejects are forbidden to be sampled by the agent. If all actions are rejected by the guardrail, the episode terminates.

Results. Figure 1 shows mean episode rewards and episode deaths under each guardrail across 10000 episodes, for different values of the rejection threshold C. The cheating guardrail achieves zero deaths for sufficiently small C, but for $C = 0.1$ its death probability is moderately high.[9] The

---

[7] Code available at `https://github.com/saifh-github/conservative-bayesian-public`.

[8] The mapping $v \mapsto \tau_v$ is not necessarily injective – multiple vectors may represent the same collection of reward distributions and therefore the same distribution over action-reward pairs.

[9] Indeed, if every action taken had a harm probability of 0.1, the probability of death across an episode would be $1 - \left((1-0.1)^{25}\right) \approx 0.93$.

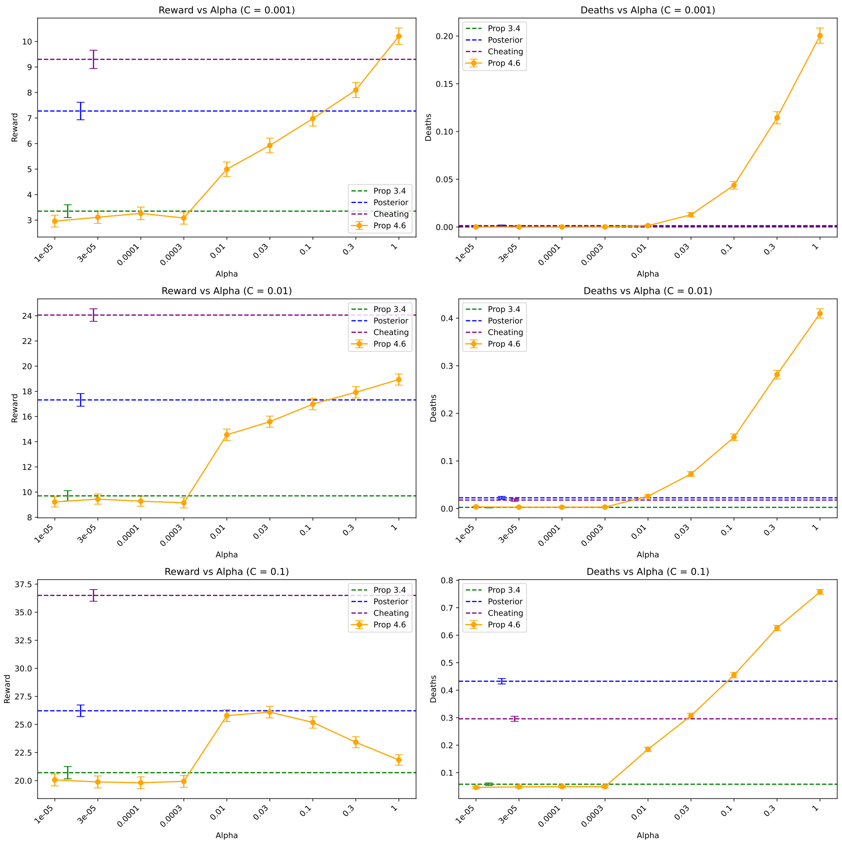
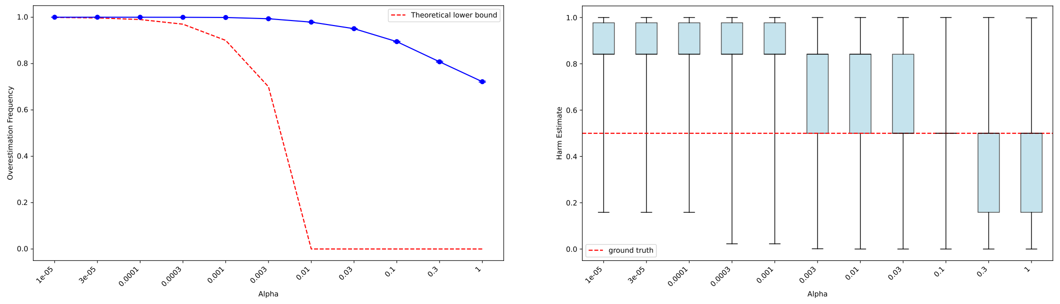


Figure 1: Mean episode deaths and reward for different guardrails in the exploding bandit setting.

posterior predictive guardrail also achieves zero deaths for small C, while for larger C it dies slightly more often and achieves somewhat lower reward compared to the cheating guardrail. The behaviour of the Proposition 4.6 guardrail depends strongly on α. When α is close to 1, actions are very rarely rejected, leading to a high probability of an early death and precluding the opportunity to obtain much reward. At the other extreme, when α is close to 0, the candidate theory set $\mathcal{M}^{\alpha}_{Z_{1:t}}$ is larger and the guardrail is extremely conservative. It rejects almost all actions, resulting in low deaths and low reward. This is the case even for larger C, since the estimated probability used to filter actions tends to overestimate an action's harm probability under the true theory. For middling values of α, the Proposition 4.6 guardrail performs similarly to the posterior predictive guardrail. The Proposition 3.4 guardrail, which makes the incorrect assumptions of i.i.d. data and distinct theories, is similarly conservative to the Proposition 4.6 guardrail with low α.

Tightness of bounds. Figure 2 takes a closer look at how often, and how tightly, the inequality in Proposition 4.6 is satisfied. For an agent following a uniform policy across 10000 bandit episodes without action rejection or death, Figure 2a shows the frequency with which $\max_{\tau \in \mathcal{M}^{\alpha}_{Z_{1:t}}} \mathrm{P}(Y_{t+1} = 1 \mid Z_{1:t}, \tau, a_{t+1})$ overestimates the true harm probability of an action. Proposition 4.6 gives us a strict lower bound of $1 - \frac{\alpha}{\mathrm{P}(\tau^*)}$ (which may be below 0) on the overestimation frequency, but the frequency significantly exceeds the bound for large α. Figure 2b shows the distribution of harm estimates for the most dangerous action, which always has a ground truth harm probability of 0.5 due to the choice of E. Note that for large α the harm of this dangerous action is usually *underestimated* – so the high overestimation rate in Figure 2a comes from actions with lower harm probabilities.



(a) The frequency with which the inequality in Proposition 4.6 is satisfied.

(b) The distribution of the right-hand side of (10), for an action with a true harm probability of 0.5.

Figure 2: Overestimate frequency and harm estimate distribution for the Proposition 4.6 guardrail for varying α.

6 Conclusion and open problems

The approach to safety verification proposed here is based on context-dependent run-time verification because the set of possible inputs for a machine learning system is generally astronomical, while the safety of the answer to a specific question is more likely to be amenable to tractable calculations. It focuses on the risk of wrongly interpreting the data, including the safety specification itself (what we called "harm" above) and exploits the fact that as more evidence is gathered (as necessarily happens with i.i.d. data) and when different theories predict different observations, the true interpretation rises towards the maximal value of the Bayesian posterior over interpretations. The bound is tighter with i.i.d. data but the i.i.d. assumption is also not realistic, and in the context of safety-critical decisions, we would prefer to err on the side of prudence and fewer assumptions. However, it provides an interesting template to think about variants of this idea in future work.

There are many open problems to consider before turning the kinds of bounds introduced above into an operational run-time safeguard:

1. **Upper-bounding overcautiousness.** Can we ensure that we do not underestimate the probability of harm but do not massively overestimate it? Some simple theories consistent with the dataset (even an arbitrarily large one) might deem non-harmful actions harmful. Can we bound how much this harm-avoidance hampers the agent? A plausible approach would be to make use of a mentor for the agent that demonstrates non-harmful behavior [5].
2. **Tractability of posterior estimation.** How can we efficiently estimate the required Bayesian posteriors? For computational tractability, a plausible answer would rely on amortized inference, which turns the difficult estimation of these posteriors into the task of training a neural net probabilistic estimator which will be fast at run-time. Recent work on amortized Bayesian inference for symbolic models, such as causal structures [9, 10], and for intractable posteriors in language models [16, 18, 36, 35, 39] – which are useful when prior knowledge is encoded in a pretrained foundation model – suggests that this is feasible. Advances in efficient and adaptive Monte Carlo methods, *e.g.*, for language models [29, 42, 23], can also be useful for this purpose, and MCMC approaches can complement and aid amortization [17, 21, 31, 20].
3. **Efficient search for a cautious theory.** How can we efficiently identify a cautious but plausible theory that upper-bounds the risk, since this requires an optimization at run-time? Again, a plausible answer is to rely on amortized probabilistic inference, *e.g.*, by sampling theories with a low or annealed temperature, a technique that has been used for decades in the nonconvex optimization literature and more recently combined with amortized sequential sampling methods [40, 41, 43, *inter alia*].
4. **Partial specification of theories.** Another issue is that a full explanatory theory τ for the evidence D will generally be too large an object to make the above bound calculations and optimization sufficiently efficient. An alternative would be to derive bounds that only require partial explanations, similarly to how humans tell cautious 'stories' about what could go wrong in uncertain circumstances when thinking about taking a potentially dangerous decision [3].
5. **Harm specification in natural language.** How is a safety specification provided in natural language, like with Constitutional AI [1], turned into a random variable over which conditional

probabilities can be estimated? Having humans label lots of situations that match the desired constraint seems impractical. The alternative proposed by Bengio [3] is to learn to interpret natural language statements as formal random variables over which probabilistic causal reasoning can occur. The ambiguity of natural language can then be captured by having a distribution over formal interpretations.

6. **Effect of approximation error.** Finally, because we envisage using machine learning to estimate the bound, we have to consider the fact that a learned predictor will be imperfect. How do we correct conservatively for the potentially dangerous errors made by this predictor? This may be particularly important if we use a neural network to estimate probabilities, since they are subject to lack of robustness and adversarial attacks [15, 38], *e.g.*, how do we make sure that the agent proposing the actions has not learned to find loopholes in the risk estimator?

References

[1] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*, 2022.

[2] Andrew Barron, Mark J. Schervish, and Larry Wasserman. The consistency of posterior distributions in nonparametric problems. *The Annals of Statistics*, 27(2):536–561, 1999.

[3] Yoshua Bengio. Towards a cautious scientist AI with convergent safety bounds, February 2024. URL `https://yoshuabengio.org/2024/02/26/towards-a-cautious-scientist-ai-with-convergent-safety-bounds/`.

[4] Yoshua Bengio, Daniel Privitera, Tamay Besiroglu, Rishi Bommasani, Stephen Casper, Yejin Choi, Danielle Goldfarb, Hoda Heidari, Leila Khalatbari, Shayne Longpre, et al. *International Scientific Report on the Safety of Advanced AI*. PhD thesis, Department for Science, Innovation and Technology, 2024.

[5] Michael K Cohen and Marcus Hutter. Pessimism about unknown unknowns inspires conservatism. *Conference on Learning Theory (COLT)*, 2020.

[6] Michael K Cohen, Marcus Hutter, and Neel Nanda. Fully general online imitation learning. *Journal of Machine Learning Research*, 23(1):15066–15095, 2022.

[7] Michael K. Cohen, Marcus Hutter, and Michael A. Osborne. Advanced artificial agents intervene in the provision of reward. *AI magazine*, 43(3):282–293, 2022.

[8] David Dalrymple, Joar Skalse, Yoshua Bengio, Stuart Russell, Max Tegmark, Sanjit Seshia, Steve Omohundro, Christian Szegedy, Ben Goldhaber, Nora Ammann, et al. Towards guaranteed safe AI: A framework for ensuring robust and reliable AI systems. *arXiv preprint arXiv:2405.06624*, 2024.

[9] Tristan Deleu, António Góis, Chris Emezue, Mansi Rankawat, Simon Lacoste-Julien, Stefan Bauer, and Yoshua Bengio. Bayesian structure learning with generative flow networks. *Uncertainty in Artificial Intelligence (UAI)*, 2022.

[10] Tristan Deleu, Mizu Nishikawa-Toomey, Jithendaraa Subramanian, Nikolay Malkin, Laurent Charlin, and Yoshua Bengio. Joint Bayesian inference of graphical structure and parameters with a single generative flow network. *Neural Information Processing Systems (NeurIPS)*, 2023.

[11] Persi Diaconis and David A. Freedman. On the consistency of Bayes estimates (with discussion). *The Annals of Statistics*, 14:1–26, 1986.

[12] J.L. Doob. Application of the theory of martingales. *Colloque International Centre Nat. Rech. Sci.*, pages 22–28, 1949.

[13] David A. Freedman. On the asymptotic behavior of Bayes' estimates in the discrete case. *The Annals of Mathematical Statistics*, 34(4):1386–1403, 1963.

[14] David A. Freedman. On the asymptotic behavior of Bayes estimates in the discrete case II. *The Annals of Mathematical Statistics*, 36(2):454–456, 1965.

[15] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *International Conference on Learning Representations (ICLR)*, 2015.

[16] Han Guo, Bowen Tan, Zhengzhong Liu, Eric P. Xing, and Zhiting Hu. Efficient (soft) Q-learning for text generation with limited good data. *arXiv preprint arXiv:2106.07704*, 2021.

[17] Edward J. Hu, Nikolay Malkin, Moksh Jain, Katie Everett, Alexandros Graikos, and Yoshua Bengio. GFlowNet-EM for learning compositional latent variable models. *International Conference on Machine Learning (ICML)*, 2023.

[18] Edward J. Hu, Moksh Jain, Eric Elmoznino, Younesse Kaddar, Guillaume Lajoie, Yoshua Bengio, and Nikolay Malkin. Amortizing intractable inference in large language models. *International Conference on Learning Representations (ICLR)*, 2024.

[19] Jacek Karwowski, Oliver Hayman, Xingjian Bai, Klaus Kiendlhofer, Charlie Griffin, and Joar Skalse. Goodhart's law in reinforcement learning. *arXiv preprint arXiv:2310.09144*, 2023.

[20] Minsu Kim, Sanghyeok Choi, Jiwoo Son, Hyeonah Kim, Jinkyoo Park, and Yoshua Bengio. Ant colony sampling with GFlowNets for combinatorial optimization. *arXiv preprint arXiv:2403.07041*, 2024.

[21] Minsu Kim, Taeyoung Yun, Emmanuel Bengio, Dinghuai Zhang, Yoshua Bengio, Sungsoo Ahn, and Jinkyoo Park. Local search GFlowNets. *International Conference on Learning Representations (ICLR)*, 2024.

[22] Victoria Krakovna, Jonathan Uesato, Vladimir Mikulik, Matthew Rahtz, Tom Everitt, Ramana Kumar, Zac Kenton, Jan Leike, and Shane Legg. Specification gaming: the flip side of AI ingenuity, 2020. URL `deepmind.com/blog/specification-gaming-the-flip-side-of-ai-ingenuity`.

[23] Alexander K Lew, Tan Zhi-Xuan, Gabriel Grand, and Vikash K Mansinghka. Sequential Monte Carlo steering of large language models using probabilistic programs. *arXiv preprint arXiv:2306.03081*, 2023.

[24] Alexander J McNeil, Rüdiger Frey, and Paul Embrechts. *Quantitative risk management: concepts, techniques and tools-revised edition*. Princeton university press, 2015.

[25] Jeffrey W. Miller. A detailed treatment of Doob's theorem. *arXiv preprint arXiv:1801.03122*, 2018.

[26] Jeffrey W. Miller. Asymptotic normality, concentration, and coverage of generalized posteriors. *Journal of Machine Learning Research*, 22(168):1–53, 2021.

[27] Alexander Pan, Kush Bhatia, and Jacob Steinhardt. The Effects of Reward Misspecification: Mapping and Mitigating Misaligned Models. In *International Conference on Learning Representations*, 2021.

[28] Richard Yuanzhe Pang, Vishakh Padmakumar, Thibault Sellam, Ankur P. Parikh, and He He. Reward Gaming in Conditional Text Generation. *arXiv preprint arXiv:2211.08714*, 2023.

[29] Du Phan, Matthew D. Hoffman, David Dohan, Sholto Douglas, Tuan Anh Le, Aaron Parisi, Pavel Sountsov, Charles Sutton, Sharad Vikram, and Rif A. Saurous. Training chain-of-thought via latent-variable inference. *Neural Information Processing Systems (NeurIPS)*, 2023.

[30] Lorraine Schwartz. On Bayes procedures. *Probability Theory and Related Fields*, 4(1):10–26, 1965.

[31] Marcin Sendera, Minsu Kim, Sarthak Mittal, Pablo Lemos, Luca Scimeca, Jarrid Rector-Brooks, Alexandre Adam, Yoshua Bengio, and Nikolay Malkin. Improved off-policy training of diffusion samplers. *arXiv preprint arXiv:2402.05098*, 2024.

[32] Joar Skalse, Lucy Farnik, Sumeet Ramesh Motwani, Erik Jenner, Adam Gleave, and Alessandro Abate. STARC: A general framework for quantifying differences between reward functions. *arXiv preprint arXiv:2309.15257*, 2024.

[33] Joar Max Viktor Skalse, Nikolaus H. R. Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward gaming. *Neural Information Processing Systems (NeurIPS)*, 2022.

[34] Joar Max Viktor Skalse, Matthew Farrugia-Roberts, Stuart Russell, Alessandro Abate, and Adam Gleave. Invariance in policy optimisation and partial identifiability in reward learning. *International Conference on Machine Learning (ICML)*, 2023.

[35] Zitao Song, Chao Yang, Chaojie Wang, Bo An, and Shuang Li. Latent logic tree extraction for event sequence explanation from LLMs. *International Conference on Machine Learning (ICML)*, 2024.

[36] Siddarth Venkatraman, Moksh Jain, Luca Scimeca, Minsu Kim, Marcin Sendera, Mohsin Hasan, Luke Rowe, Sarthak Mittal, Pablo Lemos, Emmanuel Bengio, Alexandre Adam, Jarrid Rector-Brooks, Yoshua Bengio, Glen Berseth, and Nikolay Malkin. Amortizing intractable inference in diffusion models for vision, language, and control. *arXiv preprint arXiv:2405.20971*, 2024.

[37] Jean Ville. *Étude critique de la notion de collectif*. 1939. URL http://eudml.org/doc/192893.

[38] Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does LLM safety training fail? *Neural Information Processing Systems (NeurIPS)*, 2023.

[39] Fangxu Yu, Lai Jiang, Haoqiang Kang, Shibo Hao, and Lianhui Qin. Flow of reasoning: Efficient training of LLM policy with divergent thinking. *arXiv preprint arXiv:2406.05673*, 2024.

[40] David Zhang, Corrado Rainone, Markus Peschl, and Roberto Bondesan. Robust scheduling with GFlowNets. *International Conference on Learning Representations (ICLR)*, 2023.

[41] Dinghuai Zhang, Hanjun Dai, Nikolay Malkin, Aaron Courville, Yoshua Bengio, and Ling Pan. Let the flows tell: Solving graph combinatorial problems with GFlowNets. *Neural Information Processing Systems (NeurIPS)*, 2023.

[42] Stephen Zhao, Rob Brekelmans, Alireza Makhzani, and Roger Baker Grosse. Probabilistic inference in language models via twisted sequential Monte Carlo. *International Conference on Machine Learning (ICML)*, 2024.

[43] Ming Yang Zhou, Zichao Yan, Elliot Layne, Nikolay Malkin, Dinghuai Zhang, Moksh Jain, Mathieu Blanchette, and Yoshua Bengio. PhyloGFN: Phylogenetic inference with generative flow networks. *International Conference on Learning Representations (ICLR)*, 2024.

[44] Simon Zhuang and Dylan Hadfield-Menell. Consequences of misaligned AI. *Neural Information Processing Systems (NeurIPS)*, 2020.