We thank all reviewers for their insightful suggestions. In the following, we address all the questions in order.

R1: " direct connection" - This is an interesting question! We do not think that Assumption 1 will have a direct relation in terms of Lipschitz smoothness (which is an extremely local property). For example, to infer the performance of π on M_k, no *additional* information is gained by knowing performances of π on MDPs $\{M_i\}_{i=1}^{k-2}$ when the Lipschitz constraint and the performance of a policy on MDP M_{k-1} are known. This is unlike our assumption, where data from the past MDPs $\{M_i\}_{i=1}^{k-1}$ can be informative towards inferring the performance of a policy on M_k.

"data/split", "highest lower" - We will clarify that our consideration of the lower bound for optimization was based on similar techniques used in the literature [12, 21, 52], and is not a primary contribution of our work. In Table 1, we provide an ablation study for RecoSys, for all the speeds $(0, 1, 2, 3)$. All other experimental details are the same as in Appendix E.3, except for (iv), where mean performance is optimized for instead of the lower bound. It can be seen that the safety violation rate of SPIN is robust against such hyper-parameter changes. Although, it is worth noting that too small a test-set can make it harder to pass the safety-test for executing a $\pi_c \neq \pi^{\text{safe}}$, hence performance improvement is marginally low in (i). Thank you for suggesting these experiments to improve the paper, we will include these results in the appendix.

"without striving for safety?" - If the safety check procedure for a policy's performance on a non-stationary MDP (which is one of the primary contributions of our work) is removed, then the results can be catastrophic, as can be seen in (v).

"wild bootstrap" - Time series literature is vast and it is not obvious to us which other method would be more suitable to address the challenges mentioned in Lines 147–156. A detailed discussion is provided in Appendix C.2 and C.3 regarding why several popular techniques would be ill-suited.

"lifelong", "zero-shot", "safe imitation" - Thank you for pointing these out. We will discuss these in the main paper.

"credit Assumption 1" - While we did formalize the implicit assumption made by [51] in the context of reinforcement learning, this type of assumption is popular in time series literature [6]. We will discuss this in the paper.

		train-test	0	1	2	3	0	1	2	3
(i)	SPIN	75%-25%	.56	.22	.17	.14	0.0	3.6	5.1	5.4
(ii)	SPIN	25%-75%	.48	.29	.21	.19	0.0	4.6	6.5	7.0
(iii)	SPIN (Fig. 4)	50%-50%	.62	.28	.21	.18	0.0	4.7	6.4	6.6
(iv)	SPIN-mean	50%-50%	.70	.28	.24	.19	0.2	4.9	6.3	7.1
(v)	Non-stationary + No safety	100%-0%	.73	.22	.16	.19	9.4	37.6	40.2	38.6
(vi)	Stationary + Safety (Fig. 4)	50%-50%	.85	.12	.07	.07	0.0	19.8	15.3	11.9

Table 1: (Left) Algorithm. (Middle) Improvement over π^{safe}. (Right) Safety violation percentage.

R2: Thank you for your support!

R3: "underlying linear model"- We will clarify this point of confusion in the paper. Yes, Assumption 1 requires the trend (policy's performance over time) to be a linear function of the features, ϕ, which are known ahead of time. We will state this explicitly in the paper, while reminding readers that this allows for non-linear functions when ϕ are non-linear. Additionally, we will discuss the flexibility offered by the Fourier basis for modeling a wide-class of trends [6], and emphasize Lines 271–274 to indicate that our experimental section also includes a domain (Diabetes treatment) where Assumption 1 is violated.

"explain the key differences" - We will clarify lines 57–60 to highlight that our paper extends prior work [8,51] to quantify uncertainty about a policy's future performance and to provide safety guarantees.

"conservative bandit exploration" - Thank you for pointing this out. We will include this in the main paper.

R4: "how many real world problems would satisfy these properties" - This is a good point: We should have, and will, discuss around Lines 345–347 how a practitioner can or should apply our method. Like any time-series forecasting problem, before applying our method goodness-of-fit tests [10] can be used by practitioners to check whether Assumption 1 is reasonable. For example, notice that Fig. 5 in [51] shows that this assumption is reasonable for a real digital marketing dataset. Furthermore, we will discuss how this is at least a step in the right direction: standard methods that make stationarity assumptions correspond to our method with $\phi(s) = [1]$ always (fitting a horizontal line). Even if Assumption 1 is not satisfied exactly, if the trend has an overall pattern, it is likely better to account for this overall pattern than to resort back to standard methods (fitting a horizontal line).

"algorithm would have helped" - Due to space constraints, the algorithm was deferred to Appendix D.

"pseudo samples fail", "How much data is necessary"- These are great questions! Unfortunately, there is no exact answer. Bootstrap methods provide approximate bounds and their failure rate is typically of the order $O(n^{-p/2})$, where $p \in [1, 3]$ and n is the number of samples. Lines 662–668 in the appendix provide a more detailed discussion.