We thank the reviewers for their constructive suggestions and insightful comments. We have (1) added simulations for Bayesian Causal Forests and (2) have substantially expanded the discussion in the final version to address the various reviewer comments. A summary of the added discussion is provided point-by-point below.

Reviewer 1: π is not assigned a prior. We use an empirical Bayes approach, which is, as the reviewer points out, computationally much faster than a fully hierarchical Bayesian approach of placing a prior on π. For Gaussian processes (GPs), since our approach reduces to modifying the prior covariance, the posterior can be computed using standard computational tools for GPs. Conversely, one can only sample from the hierarchical Bayes posterior using a Metropolis-Hastings-within-Gibbs-sampling algorithm, which is far slower in practice.

π requires a prior. Please note that we consider *independent* priors on (m, π, F) (for computational reasons), so that the posterior also factorizes and hence the π term is conditionally independent of ψ (see lines 94-107 and Appendix A).

Comment on Table 1. The reason for our high coverage is that our posterior bias, due to our explicit ATE bias correction (4), is (much) smaller than the posterior variance by design. As the reviewer correctly points out, asymptotic theory predicts the frequentist coverage should converge to 0.95 as the sample size increases due to the semiparametric Bernstein-von Mises theorem (cf Ray & van der Vaart (2018)). However, it is a subtle question as to when the asymptotic regime applies and our examples seem insufficiently data rich for this to be the case (e.g. $n = 1000$ observations but $d = 100$ input features).

Missing organizational section/reference for bias/language precision. We have incorporated these suggestions.

Reviewer 2: Variable selection for causal inference. We consider here a GP with squared exponential kernel with automatic relevance determination (ARD), i.e. whose data-driven lengthscale ℓ_i represents the relevance of the *ith* feature to the response surface. ARD has been used successfully for removing irrelevant inputs by several authors (see e.g. Chapter 5.1 of Rasmussen & Williams (2006)) and can thus be viewed as a form of automatic (causal) feature selection. Diagnosing missing significant covariates (confounders) is an important and difficult problem which requires further investigation in the future.

Reviewer 3: Comparison with Hahn et al. We would firstly like to clarify that our goal is to improve estimation of the *average treatment effect* (ATE) in the *presence* of heterogeneous treatment effects. It is known that naively using product priors in casual inference/missing data problems can yield biased inference [this goes back to at least Robins and Ritov (1997). The 'regularization-induced confounding' of Hahn et al. is a very nice illustration of similar ideas for the concrete and important cases of linear models and BART priors]. One solution is to reparametrize to force the missing information into the likelihood (e.g. Ritov et al. (2014), Hahn et al. (2018)), while another is to use propensity score (PS) information (Rosenbaum & Rubin (1983)).

In a nice paper Hahn et al. (2017) successfully show that this latter idea also helps Bayesian estimation using BART. Their approach is designed to improve nonparametric estimation of the *whole* response surface, which will also lead to some improvement when estimating the ATE. However, it is known that even when the prior is perfectly calibrated (i.e. all tuning parameters are set optimally) and recovers the entire response surface at the optimal rate, the posterior can still induce a bias in the *marginal posterior* for the ATE ψ that prevents efficient estimation and destroys uncertainty quantification (see e.g. Ray & van der Vaart (2018)).

The specific form in which we include the PS in our prior (4) is very deliberate - it arises as the 'least favorable direction' of the ATE in semiparametric statistical theory and is specifically designed for estimating the ATE. When either the PS or response surface are especially difficult to estimate, we expect that incorporating the PS as a covariate as in Hahn et al. (2017) will still induce a bias for the ATE (the theory in Ray & van der Vaart (2018) predicts this). In fairness, we wish to emphasize that Hahn et al.'s goal is to estimate the *entire* response surface, for which they provide excellent results, which is a different problem to estimating the ATE we consider here.

Bayesian Causal Forest (BCF) simulations have been added to the paper. In summary, for estimation BCF performs well on the synthetic data (but moderately worse than our method in both the homogeneous and heterogeneous cases) and excellently on the semi-synthetic data (moderately better than our method). For uncertainty quantification, BCF typically had the shortest credible intervals with suboptimal coverage (80-85%) on the various synthetic datasets and excellent coverage ($\sim$95%) on the semi-snythetic data.

Estimation of F. We use the widely used 'Bayesian bootstrap' (BB) since (1) it is computationally much faster (you need only one costly Cholesky decomposition to generate posterior samples of the ATE whereas with the full Dirichlet process (DP) posterior we require one per posterior draw) and (2) for moderate/large sample sizes it is very close to the true DP posterior. We do not assume that 'one has observed all possible covariates', rather that our covariate samples are representative of the population. If the observed covariates greatly differ from the underlying population distribution then indeed this will not generalize well, but then neither will any prior not involving detailed outside expert information for that particular application.