Knowledge-Based Systems 315 (2025) 113138

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Accelerated experimental design using a human–AI teaming framework

Arun Kumar A.V. [a,*], Alistair Shilton [a], Sunil Gupta [a], Shannon Ryan [a], Majid Abdolshah [b,1], Hung Le [a], Santu Rana [a], Julian Berk [a], Mahad Rashid [c,1], Svetha Venkatesh [a]

[a] Applied Artificial Intelligence Institute (A^2I^2), Deakin University, Australia
[b] Amazon, Melbourne, Australia
[c] WorkSafe, Geelong, Australia

ARTICLE INFO

Keywords:
Machine learning
Experimental design
Sample-efficiency
Gaussian process
Bayesian optimization
Human–AI teaming

ABSTRACT

In this paper we propose a human–AI teaming framework for the optimization of expensive black-box functions. Inspired by the intrinsic difficulty of extracting expert knowledge and distilling it back into AI models and by observations of human behavior in real-world experimental design, our proposed algorithm lets the human expert take the lead in the experimental process. The human expert can use their domain expertise to its full potential, while the AI plays the role of a muse, injecting novelty and searching for areas of weakness to break the human out of over-exploitation induced by cognitive entrenchment. We validate our proposed algorithm using synthetic data and with human experts performing real-world experiments.

1. Introduction

Bayesian Optimization (BO) [1] is a popular sample-efficient optimization technique to solve problems where the objective is expensive. It has been successfully applied in diverse areas [2] including material discovery [3], alloy design [4] and molecular design [5]. However, standard BO typically operates *tabula rasa*, building its model of the objective from minimal priors that do not include domain-specific information. While there has been some progress made incorporating domain-specific knowledge to accelerate BO [6–8] or transfer learning from previous experiments [9], it remains the case that there is a significant corpus of knowledge and expertise that could potentially accelerate BO even further but which remain largely untapped due to the inherent complexities involved in knowledge extraction and exploitation. In particular, this often arises from the fact that experts tend to organize their knowledge in complex schema containing concepts, attributes and relationships [10], making the elicitation of relevant expert knowledge, both quantitative and qualitative, a difficult task.

Experimental design underpins the discovery of new materials, processes and products. However, experiments are costly, the target function is unknown and the search space is unclear. To be sample-efficient, the least number of experiments must be performed. Traditionally experimental design is guided by experts who use their domain expertise and intuition to formulate an experimental design, test and iterate based on observations. Living beings from fungi [11] to ants [12] and humans [13,14] face a dilemma when they make these decisions: exploit the information they have, or explore to gather new information. How humans balance this dilemma has been studied in Daw et al. [13] – examining human choices in a multi-arm bandit problem, they showed that humans were highly skewed towards exploitation. Moreover, when the task requires specialized experts, cognitive entrenchment is heightened and the balance between expertise and flexibility swings further towards remaining in known paradigms.

To break out of this, dynamic environments of engagement are needed to force experts to incorporate new points of view [15]. For such lateral thinking to catalyze creativity, Beaney [16] has further confirmed that external stimuli are crucial. For example, using random stimuli to boost creativity has been attempted in the context of games [17]. In Sentient Sketchbook, a machine creates sketches that the human can refine, and sketches are readily created through machine learning models trained on ample data. Other approaches use machine representations to learn models of human knowledge, narrowing down options for the human to consider. Recently, Vasylenko et al. [18] constructs a variational auto-encoder from underlying patterns of chemistry based on structure/composition and then a human generated hypothesis guides possible solutions. An entirely different approach refines a target function by allowing machine learning to discover

* Corresponding author.
E-mail addresses: a.anjanapuravenkatesh@deakin.edu.au (Arun Kumar A.V.), alistair.shilton@deakin.edu.au (A. Shilton), sunil.gupta@deakin.edu.au (S. Gupta), shannon.ryan@deakin.edu.au (S. Ryan), maaajid@amazon.com (M. Abdolshah), thai.le@deakin.edu.au (H. Le), santu.rana@deakin.edu.au (S. Rana), julian.berk@deakin.edu.au (J. Berk), mahad_rashid@worksafe.vic.gov.au (M. Rashid), svetha.venkatesh@deakin.edu.au (S. Venkatesh).
[1] This work was completed during the author's affiliation with the Applied Artificial Intelligence Institute (A^2I^2), Deakin University, Australia.

https://doi.org/10.1016/j.knosys.2025.113138
Received 12 June 2024; Received in revised form 23 January 2025; Accepted 5 February 2025
Available online 8 March 2025

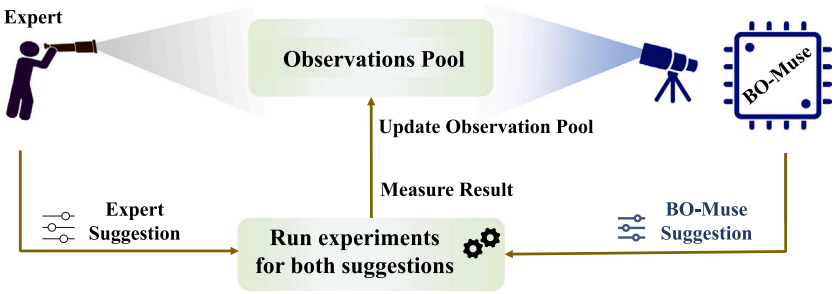

Fig. 1. Workflow of our proposed BO-Muse framework.

relations between mathematical objects, and guides humans to make new conjectures [19]. Note, however, that there is still a requirement for large datasets to formulate representations of mathematical objects, which is antithetical to sample-efficiency as typically in experimental design process we have a budget allocated on the number of experiments, data from past designs is lean, and formulation of hypothesis is difficult in this lean data space.

The use of BO for experimental design overcomes the problems of over-exploitation and cognitive entrenchment, and provides mathematically rigorous guarantees of convergence to the optimal design. However, as noted previously, this often means that domain-specific knowledge and expertise is not utilized and lost. In this paper, motivated by our aforesaid observations, we propose BO-Muse framework, which empowers human experts to lead the experimental design process. Instead of enriching AI models with expert knowledge to potentially accelerate Bayesian optimization, the proposed approach incorporates an AI "muse" that enhances the expert's intuition by providing insightful suggestions. Therefore, the AI's job is to induce dynamism that helps overcome an expert's cognitive entrenchment and go beyond the state-of-the-art solutions in novel problems [20]. In contrast, the expert leverages their extensive knowledge and domain expertise to develop state-of-the-art designs or solutions. The primary contribution of this paper lies in formalizing the integration of these two roles in a flexible framework.

The proposed BO-Muse framework integrates BO into the expert's workflow (see Fig. 1), enabling the AI's exploit/explore strategy to be adjusted based on the human expert's suggestions. This approach generates a batch of recommendations (or suggestions) from both the human expert and the AI at each iteration. The resulting batch of designs are experimentally evaluated and shared with both the AI and human experts to update their corresponding model beliefs. This process is repeated until the optimum target is achieved or the design budget is exhausted. We validate BO-Muse framework through standard optimization benchmark suites. Additionally, we collaborate with human experts to deploy and evaluate BO-Muse in complex real-world tasks. Our main contributions in this paper are as follows:

- **Design of the BO-Muse Framework:** A formal framework that enables collaboration between human experts and AI, leveraging the human's deep understanding of the problem and the AI's capability to utilize rigorous models, ultimately enhancing the sample efficiency of experimental design process.
- **Algorithm Design for Enhanced Exploration**: Development of an algorithm that mitigates the human tendency to favor exploitation by effectively enhancing AI's exploration. The proposed algorithm allows the human model to be mis-specified (i.e. allow it to deviate from the true function), which is crucial as experts rarely have perfect model.
- **Experimental Validation:** A comprehensive validation of the BO-Muse framework through standard optimization benchmark suites. We provide empirical evaluations with complex real-world tasks (for e.g., designing a protective shield for spacecrafts). As depicted in the results, BO-Muse highlights the potential benefits of the synergy between AI and human experts that can be beneficial in designing real-world novel products and processes.

2. Background

2.1. Related work on human-AI teaming

In recent years, a diverse array of human–AI teaming methodologies has been proposed, emphasizing the dynamic interplay between human experts and AI. These approaches focus on several critical factors that influence the effectiveness and efficiency of the collaboration. Some of the key factors include (i) the type and degree of involvement of each entity — whether the human or AI takes a leading role, operates as an assistant, or functions equally, (ii) varying knowledge levels of human experts and AI that bring distinct but complementary strengths, and (iii) the ultimate goal of the teaming process, which can range from problem-solving and decision-making to creative tasks or process optimization. This holistic view ensures that the methodologies are adaptable to different contexts and objectives, fostering seamless and productive human–AI teaming. For E.g., mixed initiative creative interfaces [17] in the context of creative support propose a tight coupling of human and AI to foster creativity. Deterding et al. [21] identified open challenges including "what kinds of human–AI co-creativity can we envision across and beyond creative practice?". Also of importance, though beyond the scope of this study, is the design of interfaces for such systems [22] and how the differing ways human and AI express confidence affects performance [23].

In recent years, AI systems have increasingly been utilized to assist in human decision-making, particularly in critical domains such as medicine [24,25] and judiciary [26,27], due to its exceptional capabilities. However, a key challenge is trusting AI's decision and reasoning. Fostering this trust requires aligning the AI's behavior with the human mental model. There has been an increased emphasis on incorporating human mental models into the AI modeling [28]. Green and Chen [29] proposed a gradient-boosted trees-based algorithm-in-the-loop framework to assist in predicting outcomes of financial applications and judicial pretrials. AI possess remarkable capabilities to process heaps of raw data and identify complex patterns, but they often struggle in scenarios requiring nuanced judgment and domain or ethical reasoning. In such cases, the involvement of humans is indispensable. There have been many methodologies [30,31] that let AI systems defer to humans for decision-making when it lacks confidence. [32,33] proposed to build an AI model with a query policy that calculates the prediction uncertainties of AI model online and defer to human experts for decision-making.

The literature presents various methodologies for facilitating human–AI collaboration. Some approaches [34,35] enable direct, synchronous interaction between humans and AI, while others [36–38] support asynchronous collaboration, where each party intervenes and provides inputs independently as needed. In a general human–AI teaming context, clearly defining the roles and responsibilities of each entity is

critical to understanding how they complement one another. Depending on this dynamic, human–AI teaming strategies may involve human experts learning from AI feedback, AI learning from humans, or a bi-directional learning paradigm in which both human and AI evolve together through mutual interaction and knowledge exchange.

2.1.1. Human-AI teaming based Bayesian optimization

Bayesian Optimization (BO) [39] is an efficient tool for the global optimization of unknown objective functions. In its vanilla form, BO assumes the objective function to be completely black-box, however, in many cases, domain experts provide valuable insights into the unknown objective function, that can be incorporated to boost BO performance. Further, most BO applications involve experts or skilled practitioners in the loop, creating an ideal setting for human–AI teaming.

In the BO literature, several methods have been proposed to achieve human–AI teaming by integrating expert feedback. In the early attempt, Li et al. [40] introduced a method to incorporate monotonic information about the objective function into BO by drawing inspiration from the Gaussian Process (GP) monotonicity framework [41]. There were similar efforts [42–44] to incorporate shape constraints into GPs. Andersen et al. [45] explored a BO framework tailored for uni-modal objective functions. Jauch and Peña [46] proposed a novel approach to incorporate prior knowledge about the shape of an objective function. Venkatesh et al. [8] introduced a framework to integrate trends and varying smoothness of the objective function known apriori in the optimization process to further boost its performance. Li et al. [6] developed a framework to leverage experimenters' hunches about priors. Similarly, Hvarfner et al. [7] proposed πBO to specify user prior beliefs about the location of the optimum in BO. Building on πBO, ColaBO proposed in [47] allows experts to interact with BO and incorporate user knowledge about the optimum, optimal range and preferences. These approaches aim to collaborate with experts by incorporating their static knowledge, such as shapes, trends, or priors, into the optimization process. However, in practice, human experts may not always be able to provide such explicit static knowledge prior to the optimization process.

The recent efforts [36,48–50] aimed at building collaboration platforms to harness the synergy between experts and AI. Venkatesh et al. [36] proposed a novel human–AI teaming approach where AI primarily drives the search for optimum, while intermittently incorporating feedback from experts with deep domain knowledge. Experts contribute by adjusting current recommendations or by indicating favorable and unfavorable search regions based on existing observations. Xu et al. [34] proposed a human–AI collaborative platform that lets human experts act as a black-box classifier to accept or reject the BO suggestions with low confidence. Khoshvishkaie et al. [35] proposed a human–AI collaborative platform where autonomous yet interdependent agents strategically plan to achieve a common objective, with each agent controlling a specific aspect of the decision-making process. Cissé et al. [51] proposed HypBO framework that used human hypotheses as soft constraints to achieve better-informed sampling, thereby improving convergence. Gao et al. [52] proposed a human–AI decision collaboration platform that complements human experts using a counterfactual policy optimization based algorithm.

In the recent past, there has been a surge in the interest towards explainable BO in the context of human–AI teaming. Adachi et al. [53] proposed CoExBO a Collaborative and Explainable BO to integrate human expertise directly into the BO process via a feedback loop where human preferential inputs are used to influence the model built. While CoExBO captured expert preferences directly on the objective function, BOAP proposed in [38] aims at supplementing the GP surrogate models with expert preferences about the abstract properties that are not directly measurable. Rodemann et al. [37] proposed ShapleyBO to interpret the candidate suggestions of BO. Further, ShapleyBO facilitates a human–machine interface, enabling users to intervene when the AI suggestions do not align with human intuition or reasoning. Chakraborty et al. [54] proposed TNTRules (TuneNoTune Rules), for explaining BO as a set of global rules and local actionable explanations through multi-objective optimization.

The proposed BO-Muse framework distinguishes itself from the literature by positioning the human expert to lead the optimization, with the AI serving as a supportive "muse". BO-Muse aims to mitigate cognitive entrenchment in the human expert, enhancing the overall effectiveness of the optimization.

2.2. Bayesian optimization

Bayesian Optimization (BO, Brochu et al. [39]) is an optimization method for solving the problems of the form:

$$\mathbf{x}^\star = \text{argmax}_{\mathbf{x}\in\mathbb{X}} f^\star(\mathbf{x})$$

in the least possible number of iterations, where $f^\star$ is an expensive to evaluate black-box function. Bayesian optimization models $f^\star$ as a draw from a Gaussian Process GP$(0, K)$ with prior covariance (kernel) K [55]. At each iteration t, BO recommends the next function evaluation point $\mathbf{x}_t$ by optimizing a (cheap-to-evaluate) acquisition function $a_t : \mathbb{X} \to \mathbb{R}$ based on the posterior mean and variance given a dataset of observations $\mathbb{D}_{t-1} = \{(\mathbf{x}_i, y_i) : i \in \mathbb{N}_{t-1}\}$:

$$\mu_{t-1}(\mathbf{x}) = \mathbf{y}_{t-1}^{\mathrm{T}} \left(\mathbf{K}_{t-1} + \sigma^2\mathbf{I}\right)^{-1} \mathbf{k}_{t-1}(\mathbf{x})$$
$$\sigma_{t-1}^2(\mathbf{x}) = K_{t-1}(\mathbf{x},\mathbf{x}) - \mathbf{k}_{t-1}^{\mathrm{T}}(\mathbf{x})\left(\mathbf{K}_{t-1}+\sigma^2\mathbf{I}\right)^{-1}\mathbf{k}_{t-1}(\mathbf{x})$$

where $\mathbf{y}_{t-1} = [y_i]_{i\in\mathbb{N}_{t-1}}$ is the set of observed outputs, $\mathbf{K}_{t-1} = [K(\mathbf{x}_i, \mathbf{x}_j)]_{i,j\in\mathbb{N}_{t-1}}$ and $\mathbf{k}_{t-1}(\mathbf{x}) = [K(\mathbf{x}_i,\mathbf{x})]_{i\in\mathbb{N}_{t-1}}$. Experiments evaluate $y_t = f^\star(\mathbf{x}_t) + \nu_t$, where ν_t is noise, the GP model is updated to include the new observation, and the process repeats either until a convergence criteria is met or a fixed budget of evaluations is exhausted. Several acquisition functions exist e.g. Expected Improvement (EI, [56]), GP-Upper Confidence Bound (GP-UCB) [57], Entropy search schemes [58–60], Knowledge gradient [61] etc. We use GP-UCB as it is particularly amenable for theoretical analysis. GP-UCB uses:

$$a_t(\mathbf{x}) = \mu_{t-1}(\mathbf{x}) + \beta_t^{1/2}\sigma_{t-1}(\mathbf{x}) \tag{1}$$

Here β_t is a variable controlling the trade-off between exploitation of known maxima (if β_t is small) and exploration (if β_t is large). The optimization performance depends on β_t. To achieve sub-linear convergence [62] recommend setting the value $\beta_t = \chi_t$, where:

$$\chi_t = \left(\frac{\Sigma}{\sqrt{\sigma}}\sqrt{2\ln\left(\frac{1}{\delta}\right) + 1 + \gamma_t} + \|f^\star\|_{\mathcal{H}_K}\right)^2 \tag{2}$$

and γ_t is the maximum information gain (maximum mutual information between $f^\star$ and any t observations [57]). When the assumption that $f^\star \sim \text{GP}(0, K)$ (or more generally $f^\star \in \mathcal{H}_K$, where $\mathcal{H}_K$ is the Reproducing Kernel Hilbert Space (RKHS) of the kernel K [57]) is not true then the problem is *mis-specified*. One approach to mis-specified BO is Enlarged Confidence GP-UCB (EC-GP-UCB, [63]). In EC-GP-UCB, the function closest to the objective $f^\star$ at iteration t is denoted $f_t^\# = \text{argmin}_{f\in\mathcal{H}_K}\|f^\star - f\|_\infty$, and the acquisition function is:

$$a_t(\mathbf{x}) = \mu_{t-1}(\mathbf{x}) + \left(\beta_t^{1/2} + \frac{\epsilon_t\sqrt{t}}{\sigma}\right)\sigma_{t-1}(\mathbf{x}) \tag{3}$$

where $\epsilon_t = \|f_t^\# - f^\star\|_\infty$ is the mis-specification gap. The extra term in the acquisition function is needed to ensure sub-linear convergence.

Alternatively, Rodemann and Augustin [64] investigated how imprecision in GP models affects the performance of BO, despite their flexibility. The initial sensitivity analysis in [64] concluded that the convergence of BO heavily depends on the mean function's parameters. To nullify the effects of the mis-specified prior mean and improve the convergence rates, the authors proposed Prior-mean RObust Bayesian Optimization (PROBO) using a novel acquisition function — Generalized Lower Confidence Bound (GLCB). The main idea of GLCB is to combine the posterior mean and variance of a precise GP with the upper and lower mean estimates of an imprecise GP [65].

3. Framework

Recall that our goal is to solve:

$$\mathbf{x}^\star = \text{argmax}_{\mathbf{x}\in\mathbb{X}} f^\star(\mathbf{x}), \tag{4}$$

where $f^\star : \mathbb{X} \to \mathbb{R}$ is noisy and expensive, with results (observations) $y = f^\star(\mathbf{x}) + \nu$, where ν is Σ-sub-Gaussian noise. To do this, we follow a BO methodology, running a sequence of experimental batches, indexed by $s = 1, 2, \ldots$, consisting of one human recommendation $\hat{\mathbf{x}}_s$ and one AI recommendation $\check{\mathbf{x}}_s$ (**hatted** variables for the human expert, **breved** for the AI), after which human expert and AI update their models based on the new data and the process repeats until the budget S is exhausted.

A detailed discussion on the human optimization model is provided in Appendix A. We do not assume that the human expert's model can precisely model $f^\star$, particularly in the early stages of the optimization process, due to gaps (or imperfections) in understanding the problem. Despite this imperfection, the human expert is capable of learning from observations of $f^\star$ and refining their model, thereby progressively reducing the gap between their model and reality as the optimization progresses. Therefore, we assume that the human expert maintains an implicit and evolving GP model of $f^\star$ with an unknown kernel $\hat{K}_s$. We also explicitly model $f^\star$ as a draw from a GP with prior covariance $\check{K}_s$ (the AI model), where $\check{K}_s$ may be updated after each batch. The posterior means and the posterior variances given dataset $\mathbb{D}_s = \{(\hat{\mathbf{x}}_i, \hat{y}_i), (\check{\mathbf{x}}_i, \check{y}_i) : i \le s\}$ are:

$$\begin{cases} \hat{\mu}_s(\mathbf{x}) &= \mathbf{y}_s^{\mathrm{T}}(\hat{\mathbf{K}}_s + \sigma^2\mathbf{I})^{-1}\hat{\mathbf{k}}_s(\mathbf{x}) \\ \hat{\sigma}_s^2(\mathbf{x}) &= \hat{K}_s(\mathbf{x},\mathbf{x}) - \hat{\mathbf{k}}_s^{\mathrm{T}}(\mathbf{x})(\hat{\mathbf{K}}_s + \sigma^2\mathbf{I})^{-1}\hat{\mathbf{k}}_s(\mathbf{x}) \end{cases}$$
$$\begin{cases} \check{\mu}_s(\mathbf{x}) &= \mathbf{y}_s^{\mathrm{T}}(\check{\mathbf{K}}_s + \sigma^2\mathbf{I})^{-1}\check{\mathbf{k}}_s(\mathbf{x}) \\ \check{\sigma}_s^2(\mathbf{x}) &= \check{K}_s(\mathbf{x},\mathbf{x}) - \check{\mathbf{k}}_s^{\mathrm{T}}(\mathbf{x})(\check{\mathbf{K}}_s + \sigma^2\mathbf{I})^{-1}\check{\mathbf{k}}_s(\mathbf{x}) \end{cases}$$

respectively for (implicit) human and (explicit) AI models.

We assume $f^\star \in \check{\mathcal{H}}_s = \mathcal{H}_{\check{K}_s}$ lies in the RKHS of the AI's kernel $\check{K}_s$. We do not make this assumption for the human expert, so the problem is mis-specified from their perspective. Motivated by [63], we assume that for batch s the human attempts to maximize the closest function to $f^\star$ in $\hat{\mathcal{H}}_s$:

$$\hat{f}_s^{\#} = \text{argmin}_{f\in\hat{\mathcal{H}}_s : \|f\|_{\hat{\mathcal{H}}_s} \le \hat{B}} \|f - f^\star\|_\infty$$

where $\hat{\mathcal{H}}_s = \mathcal{H}_{\hat{K}_s}$. The gap between $f^\star$ and $\hat{f}_s^{\#}$ is bounded as $\|\hat{f}_s^{\#} - f^\star\|_\infty \le \hat{\epsilon}_s$, and the human expert is able to learn (in effect, update their kernel function) to "close the gap", so $\lim_{s\to\infty}\hat{\epsilon}_s \to 0$.

A popular study [66] concluded that *"... GP-UCB show higher resemblance to human search behavior over all scores.."*. Inspired from the discussion provided in [66], the human optimization is modeled using the EC-GP-UCB [63] variant (3) of the BO algorithm where the gap ϵ_t and trade-off parameter β_t are unknown, but ϵ_t is assumed to converge towards 0. Further, EC-GP-UCB can effectively model the aforementioned imperfect human experts by accounting for the possible mis-specification via the gap mentioned in Eq. (3). Therefore, EC-GP-UCB allows us to model a human whose understanding of the function is imperfect with some gap (mis-specified) while capturing the essence of GP-UCB like human search observed in [66].

As noted in [15], the human experts typically prefer working in the vicinity of known solutions (an exploitative nature also known as cognitive entrenchment) rather than objectively exploring the design space and thus no assumptions are made regarding $\hat{\beta}_s$, but it appears likely that $\hat{\beta}_s$ will be small. Therefore we control the AI trade-off $\check{\beta}_s$ to compensate by adding much needed exploration. The AI and the human expert generates recommendations using GP-UCB and EC-GP-UCB, respectively, so:

$$\begin{aligned} \hat{\mathbf{x}}_s &= \underset{\mathbf{x}\in\mathbb{X}}{\text{argmax}}\,\hat{a}_s(\mathbf{x}) &&= \hat{\mu}_{s-1}(\mathbf{x}) + (\hat{\beta}_s^{1/2} + \tfrac{\hat{\epsilon}_s}{\sigma}\sqrt{2s})\hat{\sigma}_{s-1}(\mathbf{x}) \\ \check{\mathbf{x}}_s &= \underset{\mathbf{x}\in\mathbb{X}}{\text{argmax}}\,\check{a}_s(\mathbf{x}) &&= \check{\mu}_{s-1}(\mathbf{x}) + \check{\beta}_s^{1/2}\check{\sigma}_{s-1}(\mathbf{x}) \end{aligned} \tag{5}$$

It is convenient to specify the exploration–exploitation trade-off parameters ($\hat{\beta}_s$ and $\check{\beta}_s$) relative to the formulation mentioned in Eq. (2) [62,63]. Without loss of generality we require that $\hat{\beta}_s \in (\hat{\zeta}_{s\downarrow}\hat{\chi}_s, \hat{\zeta}_{s\uparrow}\hat{\chi}_s)$ and $\check{\beta}_s = \check{\zeta}_s\check{\chi}_s$, where $0 \le \hat{\zeta}_{s\downarrow} \le 1 \le \hat{\zeta}_{s\uparrow} \le \infty$, $\check{\zeta}_s \ge 1$ and:

$$\begin{aligned} \hat{\chi}_s &= \left(\tfrac{\Sigma}{\sqrt{\sigma}}\sqrt{2\ln(1/\delta) + 1 + \hat{\gamma}_s} + \|\hat{f}_s^{\#}\|_{\hat{\mathcal{H}}_s}\right)^2 \\ \check{\chi}_s &= \left(\tfrac{\Sigma}{\sqrt{\sigma}}\sqrt{2\ln(1/\delta) + 1 + \check{\gamma}_s} + \|f^\star\|_{\check{\mathcal{H}}_s}\right)^2 \end{aligned} \tag{6}$$

where $\hat{\gamma}_s$ and $\check{\gamma}_s$ are, respectively, the max information-gain for human expert and the AI. We show in Appendix B that this suffices to ensure sub-linear convergence for any human expert selections, and moreover that the convergence rate can be improved beyond what can be achieved by standard GP-UCB if the human operates as described here. We refer to Appendix B for a detailed theoretical analysis of `BO-Muse` framework and its sub-linear regret bound under appropriate assumptions.

`BO-Muse` described as per Eq. (5) assumes a scenario where the AI model is specified, and the human expert model is mis-specified. We cover the other possible cases where both the AI and human expert models are assumed to be mis-specified (or both specified, or even the case where the AI is mis-specified while the human expert is not) in Appendix D. Furthermore, our theoretical framework and the corresponding proofs provided in Appendix D, encompass the most general case involving arbitrarily many ($\check{p}$) AI agents and arbitrarily many ($\hat{p}$) human experts.

3.1. The BO-Muse Algorithm

The proposed `BO-Muse` method is shown in algorithm 1, where $f^\star$ is optimized using a sequence of batches $s = 1, 2, \ldots, S$, each containing one human and one AI recommendation, respectively $\hat{\mathbf{x}}_s$ and $\check{\mathbf{x}}_s$. The AI recommendation minimizes the GP-UCB acquisition function on the AI's GP posterior, with the exploration/exploitation trade-off $\check{\beta}_s$ given (see Appendix C). The human recommendation is assumed to be *implicitly* selected to minimize the EC-GP-UCB acquisition function (3), where the exploitation/exploration trade-off $\hat{\beta}_s$ is unknown but assumed to lie in the range $\hat{\beta}_s \in (\hat{\zeta}_{s\downarrow}\hat{\chi}_s, \hat{\zeta}_{s\uparrow}\hat{\chi}_s)$, and the gap $\hat{\epsilon}_s$ is unknown but assume to converge to 0 at the rate specified in Theorem 1 of Appendix B.

We use two approximations in the definition of $\check{\beta}_s$. Following the standard practice, we approximate the max information gain as:

$$\check{\gamma}_s = \sum_{(\mathbf{x},\ldots)\in\mathbb{D}_{s-1}} \ln\left(1 + \sigma^{-2}\check{\sigma}_i^2(\mathbf{x})\right) \tag{7}$$

and we approximate the RKHS norm bound $\check{B} = \|f^\star\|_{\mathcal{H}_s}$ as $\|\check{\mu}_s\|_{\mathcal{H}_s}$, which is updated using `max` to ensure it is non-decreasing with s (this will tend to over-estimate the norm, but this should not affect the algorithm's rate of convergence). Finally, for the simplicity in presentation we assume that $\sigma = \Sigma$ (the model parameter matches the noise).

Algorithm 1 `BO-Muse`

1: **Input:** Initial samples $\mathbb{D}_0 = \{(\mathbf{x}, y)\}$, AI prior GP$(0, K)$. Let $\check{B} = 1$.
2: **for** $s \in 1, 2, \ldots, S$ **do**
3: Set $\check{\beta}_s = 7\left(\sqrt{\sigma}\sqrt{2\ln(1/\delta) + 1 + \check{\gamma}_s} + \check{B}\right)^2$ as per (6), (7) and (C.3).
4: AI recommends $\check{\mathbf{x}}_s$ as $\check{\mathbf{x}}_s = \text{argmax}_{\mathbf{x}\in\mathbb{X}}\check{a}_s(\mathbf{x}) = \check{\mu}_{s-1}(\mathbf{x}) + \check{\beta}_s^{1/2}\check{\sigma}_{s-1}(\mathbf{x})$.
5: Human recommends $\hat{\mathbf{x}}_s$.
6: Run experiments to obtain $\hat{y}_s = f^\star(\hat{\mathbf{x}}_s) + \hat{\nu}_s$ and $\check{y}_s = f^\star(\check{\mathbf{x}}_s) + \check{\nu}_s$.
7: Set $\mathbb{D}_s = \mathbb{D}_{s-1} \cup \{(\hat{\mathbf{x}}_s, \hat{y}_s), (\check{\mathbf{x}}_s, \check{y}_s)\}$ and update the AI GP posterior.
8: Set $\check{B} = \max\{\check{B}, \mathbf{y}_s^{\mathrm{T}}(\mathbf{K}_s + \sigma^2\mathbf{I})^{-1}\mathbf{y}_s\}$.
9: **end for**

The primary goal of `BO-Muse` algorithm is to break the cognitive entrenchment of human experts by introducing novelty into the candidate recommendations via $\check{\mathbf{x}}_s$ (Step 4). This encourages human experts

Table 1
Synthetic Functions. Analytical forms are provided in the second column and the last column depicts the high level features used by a simulated human expert.

Functions	$f^\star(\mathbf{x})$	High level features
Matyas-2D	$0.26 * (x_1^2 + x_2^2) - 0.48 * (x_1 * x_2)$	$x'_1 = x_1^2,\ x'_2 = x_2^2$ $x'_3 = x_1 * x_2$
Ackley-4D	$-ae^{-b\sqrt{\frac{1}{d}\|\|x\|\|_2^2}} - e^{\sqrt{\frac{1}{d}\sum_i \cos(cx_i)}} + a + e^1$	$x'_1 = \cos(x_1),\ x'_2 = \cos(x_2)$ $x'_3 = \cos(x_3)$ $x'_4 = \cos(x_4),\ x'_5 = \|\|\mathbf{x}\|\|_2$
Rastrigin-5D	$10d + \sum_i^d [x_i^2 - 10\cos(2\pi x_i)]\ \forall i \in \mathbb{N}_5$	$x'_i = x_i^2\ \forall i \in \mathbb{N}_5$ $x'_{j+5} = \cos x_j\ \forall j \in \mathbb{N}_5$
Levy-6D	$\sin^2(\pi w_1) + s + (w_d - 1)^2[1 + \sin^2(2\pi w_d)]$ where $s = \sum_i^{d-1}(w_1 - 1)^2[1 + 10\sin^2(\pi w_i + 1)]$ and $w_i = 1 + \frac{x_i - 1}{4}\ \forall i \in \mathbb{N}_6$	$x'_1 = (\sin x_1)^2$, $x'_{j+1} = x_j^2(\sin x_j)^2$ $\forall j \in \mathbb{N}_6$

to explore beyond their existing biases and preconceptions, fostering creative and innovative solutions (informed suggestions $\hat{\mathbf{x}}_s$ in Step 5). Through this process, an implicit knowledge-sharing mechanism is established, where both human experts and AI systems contribute suggestions that are accumulated in a common observation pool $\mathbb{D}_s$. The shared observation pool acts as a repository of evolving knowledge. For example, over time, human experts can observe and draw inspiration from AI suggestions that may highlight unexplored regions of the design space or challenge human assumptions, thereby enriching the human expert's understanding of the design space. Conversely, the AI benefits from the human expert's domain knowledge by incorporating their suggestions into the optimization process, ensuring alignment with real-world constraints. This bidirectional flow of information between human experts and AI fosters a collaborative optimization loop, where the complementary strengths of both the entities are leveraged to achieve better teaming performance.

4. Experiments

We validate the performance of the proposed `BO-Muse` algorithm in the optimization of synthetic benchmark functions, and the real-world tasks involving human experts. In all the experiments Squared Exponential (SE) kernel is used with the associated hyper-parameters estimated using maximum-likelihood estimation. The sample-efficiency of `BO-Muse` framework and other standard baselines is measured in terms of the *simple regret* (r_t) given by:

$$r_t = f^\star(\mathbf{x}^\star) - \max_{\mathbf{x}_t \in D_t} f^\star(\mathbf{x}_t) \tag{8}$$

where $f^\star(\mathbf{x}^\star)$ is the true global optima and $f^\star(\mathbf{x}_t)$ is the best solution observed in t iterations. All the experiments were run on an Intel Xeon CPU@ 3.60 GHz workstation with 16 GB RAM capacity.

4.1. Experiments with optimization benchmark functions

We have evaluated `BO-Muse` on synthetic test functions covering a range of dimensions, as detailed in Table 1. We compare the sample-efficiency of `BO-Muse` with **(i) Generic BO:** A standard GP-UCB based BO algorithm with the exploration–exploitation trade-off factor (β) set as per [57]; **(ii) Simulated Human:** A simulated human with access to higher level properties (high level features in Table 1) that may help to model the optimization function more accurately; and **(iii) Simulated Human + PE**: A simulated human teamed with an AI agent using a pure exploration strategy (that is, an AI policy with $\breve{\zeta}_s = \infty$). To simulate a human expert (with high exploitation), a standard BO algorithm with small exploration factor maximizing $\hat{a}_s(\mathbf{x}) = \hat{\mu}_s(\mathbf{x}) + 0.001\hat{\sigma}_s(\mathbf{x})$ is used. Furthermore, we have ensured to allocate the same function evaluation budget for all the competing methods. For a d dimensional problem, we use $d + 1$ initial observations and optimize for $10 \times d$ iterations i.e., the evaluation budget for the synthetic experiments is set to $10 \times d$ function evaluations.

Fig. 2 shows the simple regret computed for Matyas-2D, Ackley-4D, Rastrigin-5D and Levy-6D functions averaged over 10 randomly initialized runs. As observed from the plots, `BO-Muse` consistently outperforms most baselines. The results with the Simulated Human baseline also show comparable performance, which we attribute to their access to crucial high-level features that significantly reduce the complexity of the search space. The slow convergence of Generic BO baseline indicates the importance and potential benefits of human–AI teaming. Interestingly, the poor performance of Simulated Human + PE is due to the over-exploration used by the pure exploration strategy which maximizes only predictive uncertainty.

Additionally, we have used COCO (COmparing Continuous Optimizers) platform [67,68] for systematic comparisons of the proposed `BO-Muse` framework against other competing algorithms. COCO platform provides a Black-box Optimization Benchmarking (BBOB) test suite containing 24 synthetic, noiseless, single objective and scalable test functions available in 2, 3, 5, 10, 20 dimensions. We refer readers to Appendix E for a detailed discussion on the test suite, experimental setup and the empirical results.

Further, we have conducted an ablation study by varying the human exploitation–exploration trade-off ($\hat{\beta}_s$) to simulate a range of over-exploitative to over-explorative experts. These ablation study results are provided in Appendix F.1. To further understand the behavior of `BO-Muse` framework with other human acquisition functions, we have also simulated an Expected Improvement (EI) acquisition function based human expert. The results of this ablation study are provided in Appendix F.2.

4.2. Real-world experiments

We now present our complex real-world optimization tasks experiments.

4.2.1. Classification tasks — support vector machines and random forests
Experimental set-up. In this experiment, our task is to choose optimal hyper-parameters for Support Vector Machine (SVM) and Random Forest (RF) classifiers operating on real-world *Biodeg* dataset from UCI repository [69]. The dataset is divided into random 80/20 train/test splits. We set up two human expert teams. Each member of Team 1 works in partnership with `BO-Muse`, whilst members of Team 2 work individually without `BO-Muse`. We recruited 8 participants[2] consisting of 4 postdoctoral researchers and 4 postgraduate research students. 2 postdocs and 2 students are allocated to each team randomly so that each team has 4 participants in total, with roughly similar expertise. Each participant is given the same evaluation budget i.e., 3 random initial designs + 30 further iterations.

[2] Necessary ethics approval obtained.

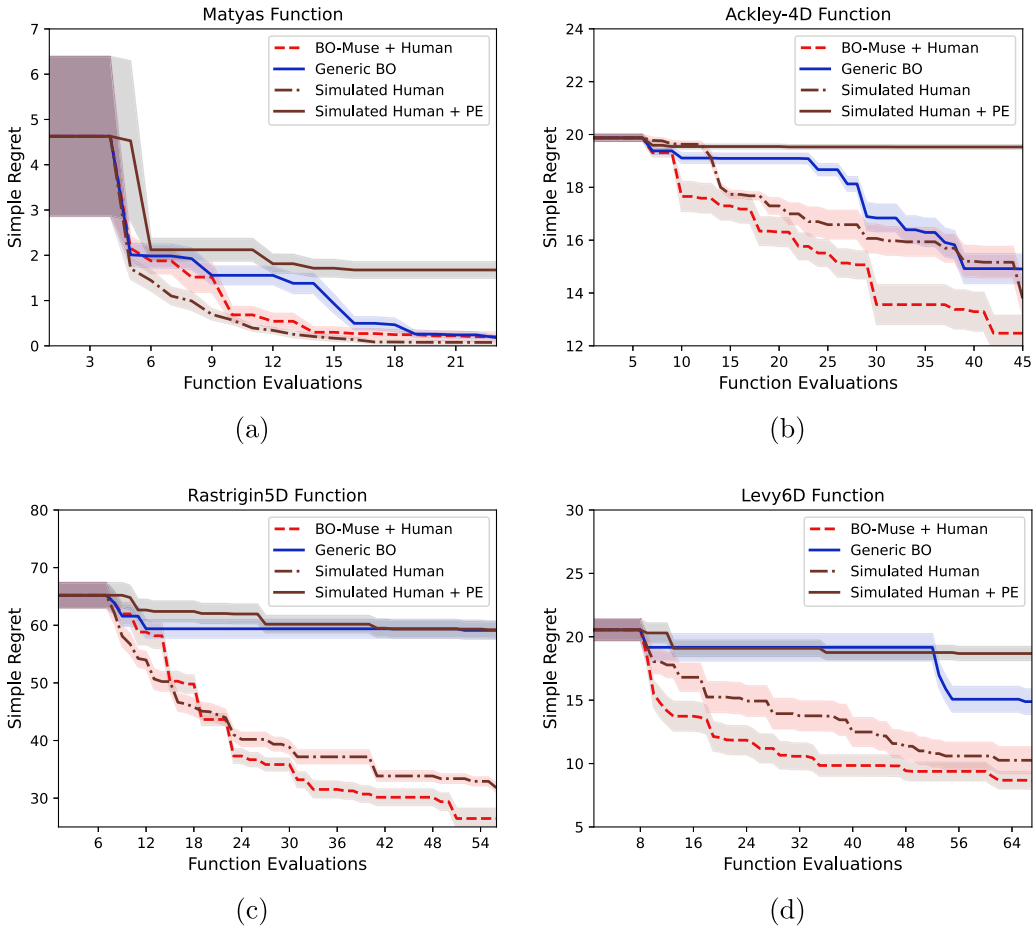

Fig. 2. Simple regret versus iterations for (a) Matyas-2D, (b) Ackley-4D, (c) Rastrigin-5D, and (d) Levy-6D functions. We plot the mean regret along with its standard error obtained after 10 repeated runs with random initialization.

Performance metric. At the end of each iteration, the test classification error for the suggested hyper-parameter set is computed. The overall performance is measured using simple regret, which will be the minimum test classification error observed so far. The individual results from each of the aforesaid teams are averaged to compare (Team 1) `BO-Muse` + Human vs (Team 2) Human alone (baseline). Further, we report the performance of Generic BO (AI alone) to demonstrate the efficacy of the approach. Further, the same seed initialization and the same evaluation budget is used for all the algorithms.

Interfacing with experts. The human experts perform two hyper-para meter tuning experiments to minimize the test classification error of SVM and RF. Each expert is provided with a simple graphical interface that shows accumulated observations of classifier performance as a function of hyper-parameters along with the best result thus far. Experts suggest the next hyper-parameter set by clicking at a point of their choice inside the plot. Same interface is provided to both teams.

Experiment 1 – SVM classification. In this classification experiment C-SVM classifiers with Radial Basis Function (RBF) kernel are considered. LibSVM [70] is used to implement SVM classifier with hyper-parameters: kernel scale γ and the cost parameter C. The SVM hyper-parameters *i.e.,* γ and C are tuned in the exponent space of $[-3, 3]$.

Experiment 2 – random forest classification. In the classification tasks with RFs we tune the hyper-parameters *maximum depth of the decision tree* and the *number of samples per split* in the range $(0, 100]$, and $(1, 50]$, respectively.

The results of our classification experiments are depicted in Figs. 3(a) and 3(b). In both experiments, the `BO-Muse` + Human team outperforms the experts working on their own and the Generic BO baseline. It is noteworthy that experts demonstrate similar optimization performance compared to Generic BO in RF hyperparameter tuning experiments. We believe this trend may be attributed to the low complexity of the search space, which allows any optimization algorithm to converge quickly.

4.2.2. Space shield design application

Our third experiment applies `BO-Muse` to a real-world applied engineering problem. We consider the design of a shield for protecting spacecraft against the impact of space debris particles by partnering with a world leading impact expert. Due to the expensive nature of this experiment, it was not feasible to do perform multiple experiments or have access to many human experts. Therefore, this design experiment was primarily done to showcase an application rather than an evaluation. The detailed experimental set-up and results are discussed in Appendix H.

In this experiment (see Table H.7 in Appendix H) we observe the human expert initially exploring solutions based on the state-of-the-art for more typical debris impact problems (which are normally simplified to spherical aluminium projectiles). The expert is observed to rapidly exploit their initial 3 designs to identify two feasible shielding solutions within the first four batch iterations (ID 4-11, marked in blue). The expert performs further exploitation of these successful designs (Result=0 i.e., Zero/No Perforation) over the next four batch iterations (ID 12–19, marked in brown) in an attempt to reduce the weight, but is unsuccessful. To this point the expert does not appear to have been influenced at all by the BO suggestions. However, in the next four batch iterations (ID 20–27, marked in green) we can observe the expert

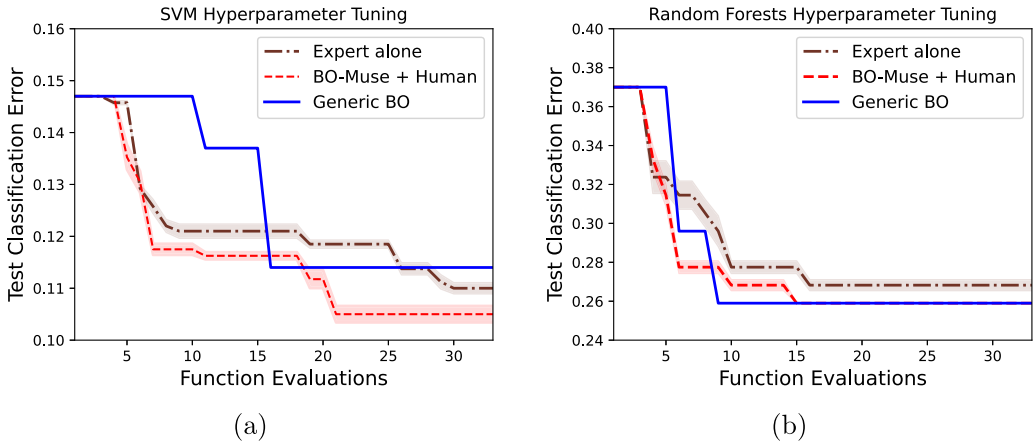

Fig. 3. Simple regret vs iterations for the hyper-parameter tuning of classifiers comparing Team 1 (BO-Muse + Human) (red) vs Team 2 (Human only) (black) vs AI alone (blue). We report the simple regret mean (along with its standard deviation) for (a) Support Vector Machines (SVM) and (b) Random Forests (RF) classifiers.

taking inspiration from the previous BO-Muse suggestions. One such exploitation results in a successful solution (ID 27), further exploitation of which provides the best solution identified by the experiment ID 29. This solution is highly *unique* for spacecraft debris shields, utilizing a polymer outer layer in contact with a metallic backing to disrupt the debris particle. Such a design is not reminiscent of any established flight hardware, see e.g., [71]. Thus, BO-Muse is demonstrated to have performed its role as hypothesized, inspiring the expert with novel designs that are subsequently exploited by the expert.

It is evident from both the synthetic and complex real-world experiments, human–AI teaming achieved via BO-Muse framework can significantly accelerate the discovery of new products and processes by overcoming an expert's cognitive entrenchment.

5. Conclusion

We have introduced an innovative framework, BO-Muse for human–AI teaming aimed at expediting expensive experimental design tasks. BO-Muse empowers the human expert to lead in the experimental process, enabling them to fully leverage their domain expertise, while the AI acts as a muse, introducing novelty and exploring search spaces that the human expert may have overlooked, effectively mitigating the over-exploitation caused by cognitive entrenchment. We designed and developed an algorithm that breaks expert's cognitive entrenchment by inducing novelty via AI "muse" suggestions. We have validated the effectiveness of BO-Muse framework through standard synthetic optimization benchmark suites. We have empirically evaluated the effectiveness of the proposed framework in accelerating real-world complex experimental design tasks.

BO-Muse addresses the issue of cognitive entrenchment in experts by effectively enhancing AI-driven exploration. While BO-Muse is based on the assumption of cognitive entrenchment discussed in the prior studies [13,15], however, this assumption may not apply in all cases. Consequently, the algorithm's sample efficiency could be lower than expected if the AI overcompensates for the expected expert over-exploitation. Further, drawing inspiration from a prominent study [66], we model human expert predictions using EC-GP-UCB [63]. However, there may be discrepancies between the expert's actions and the GP-UCB model, which can affect efficiency. Exploring the possibility of inferring or modeling expert knowledge and behavior presents an intriguing avenue for future research.

CRediT authorship contribution statement

Arun Kumar A.V.: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation. **Alistair Shilton:** Writing – review & editing, Writing – original draft, Validation, Methodology, Investigation, Formal analysis, Conceptualization. **Sunil Gupta:** Writing – review & editing, Writing – original draft, Validation, Supervision, Methodology, Investigation, Formal analysis, Conceptualization. **Shannon Ryan:** Writing – original draft, Visualization, Validation, Resources, Methodology, Investigation, Conceptualization. **Majid Abdolshah:** Visualization, Software, Methodology, Investigation, Formal analysis, Data curation. **Hung Le:** Methodology, Formal analysis. **Santu Rana:** Validation, Supervision, Methodology, Conceptualization. **Julian Berk:** Methodology, Formal analysis. **Mahad Rashid:** Visualization, Software, Methodology, Investigation, Formal analysis, Data curation. **Svetha Venkatesh:** Writing – review & editing, Writing – original draft, Validation, Supervision, Methodology, Investigation, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Alfred Deakin Prof. Svetha Venkatesh reports financial support was provided by Australian Research Council. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was partially supported by the Australian Government through the Australian Research Council's (ARC) Discovery Project funding scheme (project ID DP210102798). The views expressed herein are those of the authors and are not necessarily those of the Australian Government or Australian Research Council (ARC).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.knosys.2025.113138.

Data availability

Data will be made available on request.

References

[1] B. Shahriari, K. Swersky, Z. Wang, R.P. Adams, N. De Freitas, Taking the human out of the loop: A review of Bayesian optimization, Proc. IEEE 104 (1) (2015) 148–175.

[2] S. Greenhill, S. Rana, S. Gupta, P. Vellanki, S. Venkatesh, Bayesian optimization for adaptive experimental design: A review, IEEE Access 8 (2020) 13937–13948, http://dx.doi.org/10.1109/ACCESS.2020.2966228.

[3] C. Li, D.R. de Celis Leal, S. Rana, S. Gupta, A. Sutti, S. Greenhill, T. Slezak, M. Height, S. Venkatesh, Rapid Bayesian optimization for synthesis of short polymer fiber materials, Sci. Rep. 7 (1) (2017) 5683.

[4] M. Barnett, M. Senadeera, D. Fabijanic, K. Shamlaye, J. Joseph, S. Kada, S. Rana, S. Gupta, S. Venkatesh, A scrap-tolerant alloying concept based on high entropy alloys, Acta Mater. 200 (2020) 735–744.

[5] R. Gómez-Bombarelli, J.N. Wei, D. Duvenaud, J.M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T.D. Hirzel, R.P. Adams, A. Aspuru-Guzik, Automatic chemical design using a data-driven continuous representation of molecules, ACS Central Sci. 4 (2) (2018) 268–276.

[6] C. Li, S. Rana, S. Gupta, V. Nguyen, S. Venkatesh, A. Sutti, D. Rubin, T. Slezak, M. Height, M. Mohammed, I. Gibson, Accelerating experimental design by incorporating experimenter hunches, in: Data Mining (ICDM), 2018 IEEE International Conference on, IEEE, 2018, pp. 257–266.

[7] C. Hvarfner, D. Stoll, A. Souza, M. Lindauer, F. Hutter, L. Nardi, π-BO: Augmenting acquisition functions with user beliefs for Bayesian optimization, 2022, arXiv preprint arXiv:2204.11051.

[8] A.K.A. Venkatesh, S. Rana, C. Li, S. Gupta, A. Shilton, S. Venkatesh, Bayesian optimization for objective functions with varying smoothness, in: Australasian Joint Conference on Artificial Intelligence, 2019, pp. 460–472.

[9] A. Shilton, S. Gupta, S. Rana, S. Venkatesh, Regret Bounds for Transfer Learning in Bayesian optimization, in: Artificial Intelligence and Statistics, Florida, USA, 2017, pp. 307–315.

[10] D.M. Rousseau, Schema, promise and mutuality: The building blocks of the psychological contract, J. Occup. Organ. Psychol. 74 (4) (2001) 511–541.

[11] S.C. Watkinson, L. Boddy, K. Burton, P. Darrah, D. Eastwood, M.D. Fricker, M. Tlalka, New approaches to investigating the function of mycelial networks, Mycologist 19 (1) (2005) 11–17.

[12] S.C. Pratt, D.J. Sumpter, A tunable algorithm for collective decision-making, Proc. Natl. Acad. Sci. 103 (43) (2006) 15906–15910.

[13] N.D. Daw, J.P. O'doherty, P. Dayan, B. Seymour, R.J. Dolan, Cortical substrates for exploratory decisions in humans, Nature 441 (7095) (2006) 876–879.

[14] J.D. Cohen, S.M. McClure, A.J. Yu, Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration, Phil. Trans. R. Soc. B 362 (1481) (2007) 933–942.

[15] E. Dane, Reconsidering the trade-off between expertise and flexibility: A cognitive entrenchment perspective, Acad. Manag. Rev. 35 (4) (2010) 579–603.

[16] M. Beaney, Imagination and creativity, vol. 4, Open University Worldwide Ltd, 2005.

[17] G.N. Yannakakis, A. Liapis, C. Alexopoulos, Mixed-initiative co-creativity, 2014.

[18] A. Vasylenko, J. Gamon, B.B. Duff, V.V. Gusev, L.M. Daniels, M. Zanella, J.F. Shin, P.M. Sharp, A. Morscher, R. Chen, et al., Element selection for crystalline inorganic solid discovery guided by unsupervised machine learning of experimentally explored chemistry, Nat. Commun. 12 (1) (2021) 1–12.

[19] A. Davies, P. Veličković, L. Buesing, S. Blackwell, D. Zheng, N. Tomašev, R. Tanburn, P. Battaglia, C. Blundell, A. Juhász, et al., Advancing mathematics by guiding human intuition with AI, Nature 600 (7887) (2021) 70–74.

[20] S. Ryan, N.M. Sushma, H. Le, A.A. Kumar, J. Berk, T. Nguyen, S. Rana, S. Kandanaarachchi, S. Venkatesh, The application of machine learning in micrometeoroid and orbital debris impact protection and risk assessment for spacecraft, Int. J. Impact Eng. 181 (2023) 104727.

[21] S. Deterding, J. Hook, R. Fiebrink, M. Gillies, J. Gow, M. Akten, G. Smith, A. Liapis, K. Compton, Mixed-initiative creative interfaces, in: Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems, 2017, pp. 628–635.

[22] J. Rezwana, M.L. Maher, Designing Creative AI Partners with COFI: A Framework for Modeling Interaction in Human-AI Co-Creative Systems, ACM Trans. Comput. Hum. Interact. (2022).

[23] M. Steyvers, H. Tejeda, G. Kerrigan, P. Smyth, Bayesian modeling of human–AI complementarity, Proc. Natl. Acad. Sci. 119 (11) (2022) e2111547119.

[24] M. Nour, Z. Cömert, K. Polat, A novel medical diagnosis model for COVID-19 infection detection based on deep features and Bayesian optimization, Appl. Soft Comput. 97 (2020) 106580.

[25] F. Gou, J. Liu, C. Xiao, J. Wu, Research on artificial-intelligence-assisted medicine: a survey on medical artificial intelligence, Diagnostics 14 (14) (2024) 1472.

[26] N. Grgić-Hlača, C. Engel, K.P. Gummadi, Human decision making with machine assistance: An experiment on bailing and jailing, Proc. the ACM Human- Comput. Interact. 3 (CSCW) (2019) 1–25.

[27] T. Drakokhrust, N. Martsenko, Artificial intelligence in the modern judicial system, J. Mod. Educ. Res. 1 (2022).

[28] G. Bansal, B. Nushi, E. Kamar, W.S. Lasecki, D.S. Weld, E. Horvitz, Beyond accuracy: The role of mental models in human-AI team performance, in: Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, vol. 7, 2019, pp. 2–11.

[29] B. Green, Y. Chen, The principles and limits of algorithm-in-the-loop decision making, Proc. ACM Human- Comput. Interact. 3 (CSCW) (2019) 1–24.

[30] A. De, P. Koley, N. Ganguly, M. Gomez-Rodriguez, Regression under human assistance, in: Proceedings of the AAAI Conference on Artificial Intelligence, 34, 2020, pp. 2611–2620.

[31] D. Madras, T. Pitassi, R. Zemel, Predict responsibly: improving fairness and accuracy by learning to defer, Adv. Neural Inf. Process. Syst. 31 (2018).

[32] M. Raghu, K. Blumer, G. Corrado, J. Kleinberg, Z. Obermeyer, S. Mullainathan, The algorithmic automation problem: Prediction, triage, and human effort, 2019, arXiv preprint arXiv:1903.12220.

[33] B. Wilder, E. Horvitz, E. Kamar, Learning to complement humans, 2020, arXiv preprint arXiv:2005.00582.

[34] W. Xu, M. Adachi, C.N. Jones, M.A. Osborne, Principled Bayesian optimisation in collaboration with human experts, 2024, arXiv preprint arXiv:2410.10452.

[35] A. Khoshvishkaie, P. Mikkola, P.-A. Murena, S. Kaski, Cooperative Bayesian optimization for imperfect agents, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2023, pp. 475–490.

[36] A.K.A. Venkatesh, S. Rana, A. Shilton, S. Venkatesh, Human-AI collaborative Bayesian optimisation, in: S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, A. Oh (Eds.), in: Advances in Neural Information Processing Systems, 35, Curran Associates, Inc., 2022, pp. 16233–16245.

[37] J. Rodemann, F. Croppi, P. Arens, Y. Sale, J. Herbinger, B. Bischl, E. Hüllermeier, T. Augustin, C.J. Walsh, G. Casalicchio, Explaining Bayesian optimization by Shapley values facilitates human-AI collaboration, 2024, arXiv preprint arXiv:2403.04629.

[38] A.K.A. Venkatesh, A. Shilton, S. Gupta, S. Rana, S. Greenhill, S. Venkatesh, Enhanced Bayesian optimization via preferential modeling of abstract properties, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2024, pp. 234–250.

[39] E. Brochu, V.M. Cora, N. de Freitas, A tutorial on Bayesian optimization of expensive cost functions, with applications to active user modeling and heirarchical reinforcement learning, 2010, arXiv:1012.2599, arXiv.org.

[40] C. Li, S. Rana, S. Gupta, V. Nguyen, S. Venkatesh, Bayesian optimization with monotonicity information, in: Workshop on Bayesian optimization at Neural Information Processing Systems, NIPSW, 2017.

[41] J. Riihimäki, A. Vehtari, Gaussian processes with monotonicity information, in: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, 2010, pp. 645–652.

[42] P.J. Lenk, T. Choi, Bayesian analysis of shape-restricted functions using Gaussian process priors, Statist. Sinica (2017) 43–69.

[43] A. Pensoneault, X. Yang, X. Zhu, Nonnegativity-enforced Gaussian process regression, Theor. Appl. Mech. Lett. 10 (3) (2020) 182–187.

[44] C. Jidling, N. Wahlström, A. Wills, T.B. Schön, Linearly constrained Gaussian processes, Adv. Neural Inf. Process. Syst. 30 (2017).

[45] M.R. Andersen, E. Siivola, A. Vehtari, Bayesian optimization of unimodal functions, NIPS Work. Bayesian Optim. (2017).

[46] M. Jauch, V. Peña, Bayesian optimization with shape constraints, 2016, arXiv preprint arXiv:1612.08915.

[47] C. Hvarfner, F. Hutter, L. Nardi, A general framework for user-guided Bayesian optimization, 2023, arXiv preprint arXiv:2311.14645.

[48] T. Savage, E.A. del Rio Chanona, Human-algorithm collaborative Bayesian optimization for engineering systems, Comput. Chem. Eng. 189 (2024) 108810.

[49] K.J. Kanarik, W.T. Osowiecki, Y. Lu, D. Talukder, N. Roschewsky, S.N. Park, M. Kamon, D.M. Fried, R.A. Gottscho, Human–machine collaboration for improving semiconductor process development, Nature 616 (7958) (2023) 707–711.

[50] N. Feith, E. Rueckert, Integrating human expertise in continuous spaces: A novel interactive Bayesian optimization framework with preference expected improvement, 2024, arXiv preprint arXiv:2401.12662.

[51] A. Cissé, X. Evangelopoulos, S. Carruthers, V.V. Gusev, A.I. Cooper, HypBO: Accelerating black-box scientific experiments using experts' hypotheses, in: Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, 2024, pp. 3881–3889.

[52] R. Gao, M. Saar-Tsechansky, M. De-Arteaga, L. Han, M.K. Lee, M. Lease, Human-AI collaboration with bandit feedback, 2021, arXiv preprint arXiv:2105.10614.

[53] M. Adachi, B. Planden, D.A. Howey, K. Maundet, M.A. Osborne, S.L. Chau, Looping in the human: Collaborative and explainable Bayesian optimization, 2023, arXiv preprint arXiv:2310.17273.

[54] T. Chakraborty, C. Seifert, C. Wirth, Explainable Bayesian optimization, 2024, arXiv preprint arXiv:2401.13334.

[55] C.E. Rasmussen, C.K.I. Williams, Gaussian Processes for Machine Learning, MIT Press, 2006.

[56] D.R. Jones, M. Schonlau, W.J. Welch, Efficient global optimization of expensive black-box functions, J. Global Optim. 13 (4) (1998) 455–492.

[57] N. Srinivas, A. Krause, S.M. Kakade, M.W. Seeger, Information-theoretic regret bounds for Gaussian process optimization in the bandit setting, IEEE Trans. Inf. Theory 58 (5) (2012) 3250–3265.

[58] P. Hennig, C.J. Schuler, Entropy search for information-efficient global optimization, J. Mach. Learn. Res. 13 (6) (2012).
[59] J.M. Hernández-Lobato, M.W. Hoffman, Z. Ghahramani, Predictive entropy search for efficient global optimization of black-box functions, Adv. Neural Inf. Process. Syst. 27 (2014).
[60] Z. Wang, S. Jegelka, Max-value entropy search for efficient Bayesian optimization, in: International Conference on Machine Learning, PMLR, 2017, pp. 3627–3635.
[61] P. Frazier, W. Powell, S. Dayanik, The knowledge-gradient policy for correlated normal beliefs, INFORMS J. Comput. 21 (4) (2009) 599–613.
[62] S.R. Chowdhury, A. Gopalan, On kernelized multi-armed bandits, in: D. Precup, Y.W. Teh (Eds.), Proceedings of the 34th International Conference on Machine Learning, in: Proceedings of Machine Learning Research, vol. 70, PMLR, International Convention Centre, Sydney, Australia, 2017, pp. 844–853.
[63] I. Bogunovic, A. Krause, Misspecified Gaussian Process Bandit Optimization, in: Advances in Neural Information Processing Systems, vol. 34, 2021, pp. 3004–3015.
[64] J. Rodemann, T. Augustin, Accounting for Gaussian process imprecision in Bayesian optimization, in: International Symposium on Integrated Uncertainty in Knowledge Modelling and Decision Making, Springer, 2022, pp. 92–104.
[65] F. Mangili, A prior near-ignorance Gaussian process model for nonparametric regression, Internat. J. Approx. Reason. 78 (2016) 153–171.
[66] A. Borji, L. Itti, Bayesian optimization explains human active search, Adv. Neural Inf. Process. Syst. 26 (2013).
[67] N. Hansen, A. Auger, R. Ros, O. Mersmann, T. Tušar, D. Brockhoff, COCO: A platform for comparing continuous optimizers in a black-box setting, Optim. Methods Softw. 36 (2021) 114–144, http://dx.doi.org/10.1080/10556788.2020.1808977.
[68] R. Le Riche, V. Picheny, Revisiting Bayesian optimization in the light of the COCO benchmark, Struct. Multidiscip. Optim. 64 (5) (2021) 3063–3087.
[69] D. Dua, C. Graff, UCI machine learning repository, 2017, URL http://archive.ics.uci.edu/ml. (Online; Accessed 22 January 2025).
[70] C.-C. Chang, C.-J. Lin, LIBSVM: a library for support vector machines, ACM Trans. Intell. Syst. Technol. (TIST) 2 (3) (2011) 1–27.
[71] E. Christiansen, J. Arnold, A. Davis, J. Hyde, D. Lear, J.-C. Liou, F. Lyons, T. Prior, M. Ratliff, S. Ryan, F. Giovane, B. Corsaro, G. Studor, Handbook for Designing MMOD Protection, Technical Report NASA/TM-2009-214785, NASA Johnson Space Center, 2009.