Explaining Bayesian Optimization by Shapley Values Facilitates Human-AI Collaboration For Exosuit Personalization

Julian Rodemann[1,2] (✉), Federico Croppi[2], Philipp Arens[3], Yusuf Sale[4,5], Julia Herbinger[6], Bernd Bischl[2,5], Eyke Hüllermeier[4,5,7], Thomas Augustin[2], Conor J. Walsh[3,8], and Giuseppe Casalicchio[2,5]

[1] CISPA Helmholtz Center for Information Security, Saarbrücken, Germany
[2] Department of Statistics, Ludwig-Maximilians-Universität (LMU), Munich, Germany
[3] John A. Paulson Harvard School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts, USA
[4] Institute of Informatics, Ludwig-Maximilians-Universität (LMU), Munich, Germany
[5] Munich Center for Machine Learning (MCML), Germany
[6] Leibniz Institute for Prevention Research and Epidemiology, Bremen, Germany
[7] German Research Centre for Artificial Intelligence (DFKI), Kaiserslautern, Germany
[8] Wyss Institute for Biologically Inspired Engineering, Harvard University, Cambridge, Massachusetts, USA

Abstract. Bayesian optimization (BO) has become indispensable for black box optimization. However, BO is often considered a black box itself, lacking transparency in the rationale behind proposed parameters. This is particularly relevant in human-in-the-loop applications like personalization of wearable robotic devices. We address BO's opacity by proposing `ShapleyBO`, a framework for interpreting BO proposals by game-theoretic Shapley values. Our approach quantifies the contribution of each parameter to BO's acquisition function (AF). By leveraging the linearity of Shapley values, `ShapleyBO` can identify the influence of each parameter on BO's exploration and exploitation behaviors. Our method gives rise to a `ShapleyBO`-assisted human-machine interface (HMI), allowing users to interfere with BO in case proposals do not align with human reasoning. We demonstrate these HMI's benefits for the use case of personalizing wearable robotic devices (assistive back exosuits) by human-in-the-loop BO. Results suggest that human-BO teams with access to `ShapleyBO` outperform teams without access to `ShapleyBO`.[9][10]

Keywords: Bayesian Optimization · Explainable AI · Interpretable Machine Learning · Shapley Values · Robotics · Human-AI Collaboration.

[9] **Open Science:** `ShapleyBO` as well as code and data to reproduce findings available at `https://github.com/rodemann/ShapleyBO`.

[10] This work builds upon the master's thesis of the second author supervised by the last author [18], and substantially extends and formalizes the results presented therein.

1 Introduction

In artificial intelligence (AI) and machine learning (ML), the black-box nature of increasingly complex models poses serious challenges to end-users and researchers alike. The terms explainable AI (XAI) and interpretable machine learning (IML) – often used interchangeably – describe efforts to help illuminate decision-making processes of ML algorithms, see [6,10] for an overview of this emerging field. While the interpretability of ML models has been extensively studied, less attention has been given to the explanation and interpretation of optimization methods, which, given their frequent use in decision making problems, may benefit particularly from increased transparency.

This paper expands the focus of interpretability to Bayesian optimization (BO) with Gaussian Processes (GPs), an optimization method, frequently used in black-box applications such as hyperparameter optimization of ML models, the sequential design of expensive computer simulations or, real-word experiments, for which gradients are difficult to compute. However, BO algorithms are often perceived as black boxes themselves. Understanding and interpreting such optimizers can increase trust in domains such as human-AI interaction, mitigating the risk for algorithmic aversion [19,13]. In addition, we will show that IML techniques can help accelerate the optimization in collaborative setups between humans and AI. Here, a human can intervene by rejecting or rectifying the proposals made by BO [48]. A better understanding of the algorithm fosters more efficient human-machine interaction, which is key in such applications, as we demonstrate in this work.

We present a method to interpret the BO's proposed parameter configurations through Shapley values, a concept from cooperative game theory that has gained much popularity in IML. Our framework `ShapleyBO` informs users about how much each parameter contributed to the configurations proposed by a BO algorithm. The key idea is to quantify each parameters' contribution to the Acquisition Function (AF) – rather than to the model's predictions, as is customary in IML [28]. Loosely speaking, the AF describes how "attractive" BO considers a given parameter configuration. A Shapley value can thus inform the user how much a single parameter contributes to this attractiveness. Since Shapley values are linear in the contributions they explain, see Axiom 3 in Section 2, they can be used to inform us about how much each parameter contributed to each component of any additive AF, such as the popular confidence bound, which is a weighted sum of the predicted mean and standard error. AFs play a critical role in BO as they define a decision metric based on which the optimization proceeds to the following iteration. In forming this decision metric, AFs balance exploration of regions with high uncertainty and exploitation of regions with high expected reward. Efficiently managing this exploration-exploitation trade-off is a central objective behind BO.

Bayesian Optimization has become particularly appealing for applications in which objective function samples are costly to obtain. A prominent example is Human-in-the-Loop (HIL) optimization to customize assistance settings for wearable robotic or prosthetic devices [54,5,20]. The goal of such HIL experi-

ments is to find a set of control parameters that maximize the efficacy of the provided assistance. This efficacy is often evaluated through physiological performance metrics such as (reductions in) metabolic demand or muscles dynamics and more recently expanded to subjective metrics such as user preference. Common to all, it is typically unclear to the user (or researcher), why a certain new parameter combination was chosen by the BO algorithm.

To address this, we propose a Human-Machine Interface (HMI) that allows users to better understand BO's proposals and utilize this understanding in deciding whether to intervene and rectify proposals if they seem undesirable for some reason, e.g., because they do not align with human preferences. Experiments on data from a real-world use case, personalizing assistance parameters for a wearable back exosuit [42,5], suggest that such an understanding can indeed help to intervene more efficiently than without the availability of Shapley values.

We summarize our contributions as follows.

(1) We explain why parameters are proposed in BO by quantifying each parameters' contribution to a proposal through Shapley values.

(2) We further distinguish between parameters that drive exploitation (mean optimization) and exploration (uncertainty reduction) in BO, utilizing the linearity of Shapley values.

(3) Exploratory uncertainty reduction is in turn disentangled into aleatoric uncertainty on the one hand and different epistemic sources of uncertainty on the other hand, which fosters theoretical understanding of BO.

(4) We test `ShapleyBO` on both noisy and noise-free optimization problems and illustrate its practical benefits, see Section 5.

(5) To compute the Shapley values, we adopt a traditional MC sampling strategy, supplemented by a novel algorithm, designed to accurately determine an adequate sample size for Shapley value estimation in BO contexts. This increases the computational efficiency of `ShapleyBO`, see supplementary material.

(6) We apply our `ShapleyBO`-based HMI to exosuit customization through human-in-the-loop BO and demonstrate that our method can speed up the procedure through more efficient HMI in a simulation study, see Section 6.

2 Background

Bayesian Optimization: BO is a popular derivative-free optimizer for functions that are expensive to evaluate and lack an analytical description. Its origin dates back to [30]. Modern use cases of BO cover engineering, drug discovery and finance as well as hyperparameter optimization and neural architecture search in ML, see e.g. [22,34,43]. BO approximates the target function through a surrogate model (SM). In the case of real-valued parameters, the SM typically is a GP. BO then combines the GP's mean and standard error predictions to construct an AF, which is then optimized to propose new points. Algorithm 1 summarizes BO applied to the problem of minimizing (w.l.o.g.) an unknown ("black-box")

objective function[11] $\Psi : \boldsymbol{\Theta} \to \mathbb{R}, \boldsymbol{\theta} \mapsto \Psi(\boldsymbol{\theta})$, where $\boldsymbol{\Theta}$ is a p-dimensional decision (parameter) space. In the human-in-the-loop setup, a user can intervene by either rejecting a proposal (line 4 in Algorithm 1) or an update (line 6 in Algorithm 1) or by proposing another configuration (line 6 in Algorithm 2), see Section 6 for details.

Algorithm 1 Bayesian Optimization

1: create an initial design $D = \{(\boldsymbol{\theta}^{(i)}, \Psi^{(i)})\}_{i=1,\dots,n_{init}}$
2: **while** termination criterion is not fulfilled **do**
3: **train** SM on data D
4: **propose** $\boldsymbol{\theta}^{new}$ that optimizes $AF(SM(\boldsymbol{\theta}))$
5: **evaluate** Ψ on $\boldsymbol{\theta}^{new}$
6: **update** $D \leftarrow D \cup (\boldsymbol{\theta}^{new}, \Psi(\boldsymbol{\theta}^{new}))$
7: **end while**
8: **return** $\arg\min_{\boldsymbol{\theta} \in D} \Psi(\boldsymbol{\theta})$

Shapley Values: Shapley values are a concept from cooperative game theory, originally introduced by [41], that can be used to measure the contribution of each feature to an ML model prediction [45]. The key idea is to consider each feature as a player in a game where the prediction is the game's payoff, and to distribute this payoff fairly among the players according to their marginal contributions. Shapley values have several desirable properties that make them appealing for interpreting optimization problems. In general, given a set of players $P = \{1, \dots, p\}$ and a value or payout function $v : 2^P \to \mathbb{R}$ that assigns a value $v(S)$ to every subset (called a *coalition* in game theory) $S \subseteq P$ (such that $v(\emptyset) = 0$), the Shapley value $\phi_j(v)$ of player j is defined as the weighted average of their marginal contributions across all possible coalitions [32]:

$$\phi_j(v) = \sum_{S \subseteq P \setminus \{j\}} \frac{|S|!\,(p - 1 - |S|)!}{p!} \left[v(S \cup j) - v(S)\right] \tag{1}$$

The Shapley value can be justified axiomatically through the properties of dummy player, efficiency, linearity, and symmetry.

- *Dummy Player:* If $v(S \cup \{j\}) = v(S)$ for player j and $\forall S \subseteq P \setminus \{j\}$, then $\phi_j(v) = 0$
- *Efficiency:* $\sum_{j=1}^{p} \phi_j(v) = v(P) - v(\emptyset)$
- *Linearity:* Given two games (P, v_1) and (P, v_2) and any $a, b \in \mathbb{R}$, the following holds: $\phi_j(av_1 + bv_2) = a\phi_j(v_1) + b\phi_j(v_2)$
- *Symmetry:* If $v(S \cup \{j\}) = v(S \cup \{l\})$ for players j, l and every $S \subseteq P \setminus \{j, l\}$, then $\phi_j(v) = \phi_l(v)$

[11] Also referred to as *target* function.

The payout function does not require any specific properties and the Shapley value can hence be used in many different applications [23, p.3]. We will particularly rely on the linearity of Shapley values when applying them to AFs in BO, see section 4.

Compared to other IML methods, such as the permutation feature importance [12,21] or the partial dependence plot [24], Shapley values have the main advantage of fairly distributing feature interactions among all involved features to quantify feature contributions. While the features become the players, the payout function is typically set to the expected output of the predictive model conditioned on the values of the features in a coalition, see [28,45] or [1, Equation 2]. Formally, let $\hat{f} : \boldsymbol{\Theta} \to \mathbb{R}$ be a prediction model on feature space $\boldsymbol{\Theta}$ and $\tilde{\boldsymbol{\theta}} \in \boldsymbol{\Theta}$ the instance to explain. Then the worth of a coalition of features $S \subseteq \boldsymbol{\Theta}$ is given by $v(S) = \mathbb{E}[\hat{f}(\boldsymbol{\theta})|\boldsymbol{\theta}_S = \tilde{\boldsymbol{\theta}}_S]$, where $\boldsymbol{\theta}_S, \tilde{\boldsymbol{\theta}}_S \in S$ are the feature vectors $\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}$ projected onto S.

3 Related Work

As mentioned in Section 2, there are only quite mild regularity conditions for a function to be explainable by Shapley values. Consequently, there exists a broad body of research on deploying Shapley values beyond classical prediction functions. Examples comprise the explanation of predictive uncertainty [50] or anomaly detection [46]. There is some work on Shapley-based explanations of optimization algorithms such as evolutionary algorithms [49] or differentiable architecture search (DARTS) in deep learning [52]. There are even efforts to utilize Shapley values to improve optimizers similar to our Shapley-assisted human-BO team. For instance, [51] solves fuzzy optimization problems by integrating Shapley values with evolutionary algortihms and [9] use Shapley values to speed up multi-objective particle swarm optimization grey wolf optimization (PSOGWO).

Generally, there has been a lot of interest in how to incorporate human knowledge in optimization loops recently [7,4,48,2,53,8] and what role IML can play in this regard [31]. This growing interest is not only sparked by fine-tuning large language models through reinforcement learning from human feedback [35], but also by chemical applications [17]. The method we apply to exosuit personalization partly builds on [18], who proposed to interpret BO by Shapley values first. Very recently, Adachi et al. [3] introduced Collaborative and Explainable Bayesian Optimization (CoExBo), building on GP-SHAP (Shapley Values explaining Gaussian Processes) [15], for lithium-ion battery design, a framework that integrates human knowledge into BO via preference learning and explains its proposals by Shapley values. Contrary to our approach, CoExBo first aligns human knowledge with BO by preference learning. In a second step, it then proposes several points and allows the user to select among them based on additionally provided Shapley values, while our Shapley-assisted human-BO team directly uses Shapley values to align a single BO proposal with human proposals.

[14] recently proposed TNTRules, a post-hoc rule-based explanation method of BO. TNTRules finds (through clustering algorithms) subspaces of the param-

eter space that should be tuned by the user. Similar to our work, it emphasizes the benefits of XAI methods in human-collaborative BO. Contrary to our work, it is a post-hoc method (`ShapleyBO` works online) and focuses on explaining the whole parameter space rather than single BO proposals.

4 Explaining Bayesian Optimization via Shapley Values

In this section, we introduce `ShapleyBO`[10] that allows to interpret BO proposals by Shapley values. Transitioning from ML models to AFs, the utilization of the Shapley value becomes remarkably straightforward, as an AF essentially represents a transformed version of a surrogate model's prediction function. Consequently, the Shapley value can be employed with any AF to evaluate the contribution of selected parameter values. Among an array of AF options, the confidence bound appears particularly suited for our approach, due to its intuitive functional form and additive nature. The **confidence bound (CB)** of a parameter vector $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ is defined as

$$cb(\boldsymbol{\theta}) = \hat{\mu}(\boldsymbol{\theta}) - \lambda\hat{\sigma}(\boldsymbol{\theta}), \tag{2}$$

where $\hat{\mu}$ and $\hat{\sigma}$ are mean and standard error estimates by the SM (here: GP), respectively; $\lambda > 0$ is a hyperparameter controlling the exploration-exploitation trade-off. The rationale behind the confidence bound is fairly intuitive: a point is deemed desirable if either (i) the mean prediction $\hat{\mu}$ is low (indicating an anticipation of a low target value, thus *exploiting* existing knowledge) or (ii) the uncertainty prediction $\hat{\sigma}$ is high (indicating limited information about the target function in that area, thus *exploring* this region of the parameter space). Proposing new samples boils down to optimizing this confidence bound. To this end, let minimizing (w.l.o.g.) the confidence bound be a cooperative game along the lines of Section 2. It shall be defined as (P, cb), or as two separate games $(P, \hat{\mu})$ and $(P, \hat{\sigma})$, with P being the grand coalition of parameters and $\hat{\mu}$ and $\hat{\sigma}$ the payout functions, respectively. According to the linearity axiom, the cb contribution of any parameter j of the parameter vector $\tilde{\boldsymbol{\theta}}$ to be explained can then be decomposed into the mean contribution $\phi_j(\hat{\mu})$ and the uncertainty contribution $\phi_j(\hat{\sigma})$:

$$\phi_j(cb) = \phi_j(\hat{\mu} - \lambda\hat{\sigma}) = \phi_j(\hat{\mu}) - \lambda\phi_j(\hat{\sigma}) \tag{3}$$

Thus, we can not only evaluate the overall contribution of each parameter θ_j, but also examine how both contributions $\phi(\hat{\mu})$ and $\phi(\hat{\sigma})$ impact and drive the selection of proposed parameter values, shedding some light on the exploration-exploitation trade-off. On the background of recent work on uncertainty quantification [26,33,27], we further aim at disentangling the uncertainty contribution $\phi_j(\hat{\sigma})$ of a parameter θ_j into its epistemic (reducible) and aleatoric (irreducible) part. Aleatoric uncertainty is typically caused by noise. This is particularly relevant in BO if the noise is heteroscedastic, i.e., dependent on $\boldsymbol{\theta}$, since decision makers are often risk-averse. In other words, when deciding

among two parameter configurations with equal mean target, most humans tend to opt for the one with lower variation. This motivates a risk-averse optimization problem: $\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} f(\boldsymbol{\theta}) - \alpha \cdot \epsilon(\boldsymbol{\theta})$ with $\epsilon(\boldsymbol{\theta})$ some noise that is non-constant in $\boldsymbol{\theta}$ and α the degree of risk-aversion. [29] propose risk-averse heteroscedastic Bayesian optimization (RAHBO) which entails minimizing (w.l.o.g.) the **risk-averse confidence bound (racb)**:

$$racb(\boldsymbol{\theta}) = \widehat{\mu}(\boldsymbol{\theta}) - \tau \cdot \widehat{\sigma}(\boldsymbol{\theta}) + \alpha \cdot \widehat{\epsilon}(\boldsymbol{\theta}), \tag{4}$$

where $\widehat{\epsilon}(\boldsymbol{\theta})$ is an on-the-fly estimate of the noise. Due to *racb*'s additive structure, `ShapleyBO` can identify each parameter's contribution to epistemic uncertainty *reduction* through $-\tau \cdot \widehat{\sigma}(\boldsymbol{\theta})$ and to aleatoric uncertainty *avoidance* through $\alpha \cdot \widehat{\epsilon}(\boldsymbol{\theta})$. By filtering out these exploratory contributions, the remainder of a parameter's overall Shapley value can be identified as the parameter's contribution to mean optimization through $\widehat{\mu}(\boldsymbol{\theta})$ (exploitation).

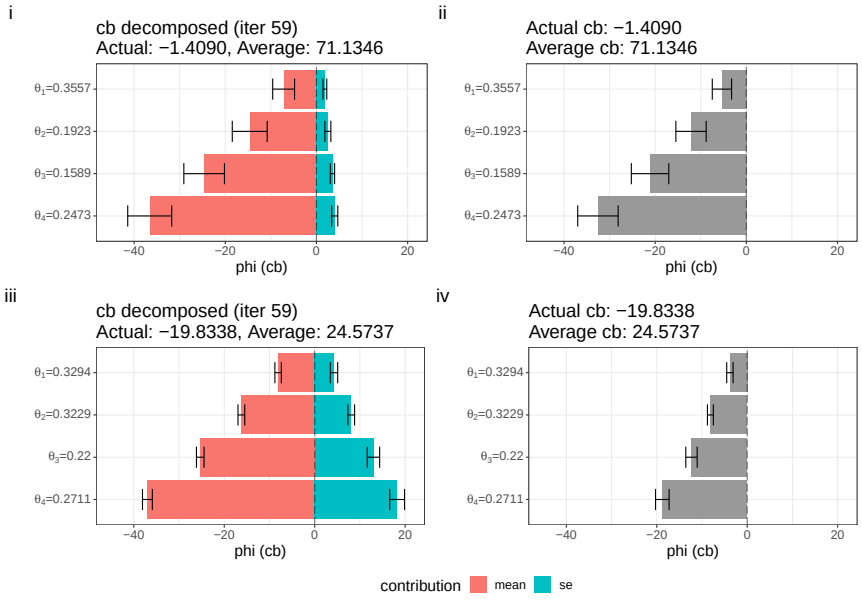

Fig. 1. `ShapleyBO` results in iteration 59 of BO on $f(\boldsymbol{\theta})$. Plots i and ii for $\lambda = 1$ and plots iii and iv for $\lambda = 10$. Contributions (*phi*) are averaged over 30 restarts for each λ. On the right, the overall contribution of the parameters is displayed (*cb* contributions), and on the left the decomposition into $\hat{\mu}$ (red, "mean") and $\hat{\sigma}$ (blue, "se") contributions. Recall that cb is minimized. Vertical axis includes the average distance of the proposed configuration from their optimum for a better interpretation. Error bars show one standard deviation. Actual: *cb* of actually proposed point. Average: Mean *cb* over all parameters.

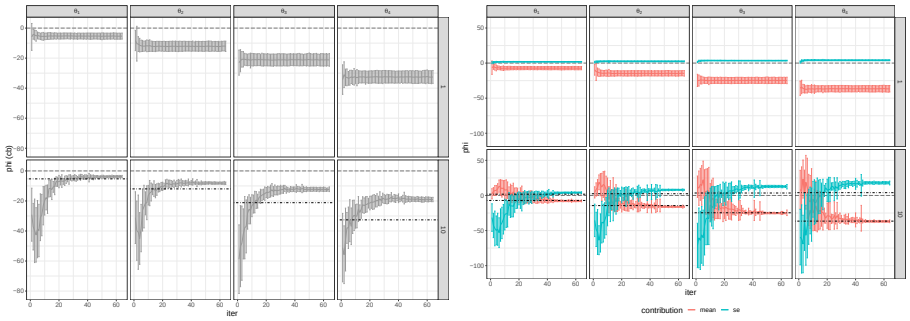

Fig. 2. Contributions curves for hyper ellipsoid optimization. Plot on top displays *cb* contributions for parameters (vertical) and λ (horizontal); beneath its decomposition into $\hat{\mu}$ (red, "mean") and $\hat{\sigma}$ (blue, "se") contributions, averaged over 30 restarts, error bars show one standard deviation. The black dot-dashed line in the $\lambda = 10$ plots displays the average contribution of the parameters in the $\lambda = 1$ run.

5 Experimental Validation

A deployment on synthetic functions allows us to validate our method, because we can formulate concrete expectations for the contributions based on the known functional form of the synthetic target function. We select a hyper-ellipsoid function, where the parameters' partial derivatives grow in j. Thus, we expect the Shapley values of parameters in BO to be higher the higher their j.

Hyper-Ellipsoid Function: Firstly, we select a hyper-ellipsoid function (Equation 5), where the partial derivatives of the parameters grow with j. Thus, we expect `ShapleyBO` to identify parameters with higher j as more influential in BO. We illustrate `ShapleyBO` by optimizing

$$f : [-5.12, 5.12]^4 \to \mathbb{R}_0^+; \boldsymbol{\theta} \mapsto f(\boldsymbol{\theta}) = \sum_{j=1}^{4} j \cdot \theta_j^2 \tag{5}$$

where f is separable and strictly convex with a unique minimizer $\boldsymbol{\theta}^* = (0, 0, 0, 0)^T$ with $f(\boldsymbol{\theta}^*) = 0$. To control for the stochastic behavior of BO, 30 repetitions of the optimization process with a bugdet of 60 function evaluations are run. Results in each iteration are then averaged over all repetitions.

As expected in light of the partial derivatives of $f(\boldsymbol{\theta})$, the contribution of θ_j grows with j, see Figures 1 and 2. In contrast, uncertainty exhibits a diminutive and adverse effect. The uncertainty measurement $\hat{\sigma}$ for the recommended setup falls below the average, thus yielding a positive payout (negative contributions). Opting for a setup with an uncertainty estimate beneath the average is deemed a strategic compromise towards enhancing mean values at the expense of exploration. `ShapleyBO` facilitates a nuanced allocation of this trade-off across parameters, see both Figures 1 and 2. We also study how the contributions change in the course of the optimization. Respective contribution paths are shown in Figure 2. Throughout the optimization process, the emphasis shifts from reducing

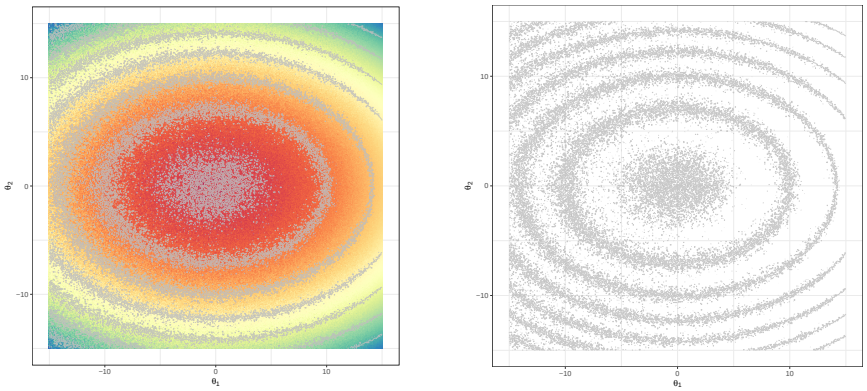

Fig. 3. Contour plots of noisy ellipsoid function $g(\boldsymbol{\theta}) + \boldsymbol{\epsilon}(\boldsymbol{\theta})$, see Equations 6 and 7. Red: low values of $g(\boldsymbol{\theta}) + \boldsymbol{\epsilon}(\boldsymbol{\theta})$; blue: high values of $g(\boldsymbol{\theta}) + \boldsymbol{\epsilon}(\boldsymbol{\theta})$. It becomes evident that the noise $\epsilon(\boldsymbol{\theta})$ varies more w.r.t. θ_1, while $g(\theta)$ is stronger affected by θ_2.

iteration	j	$\phi_j(\hat{\mu})$	$\phi_j(\hat{\sigma})$	$\phi_j(\hat{\epsilon})$	$\phi_j(racb)$
1	θ_1	-100.2	2.4	-13.9	-111.8
1	θ_2	-163.1	2.2	1.5	-159.4
2	θ_1	-87.8	2.4	-37.7	-123.1
2	θ_2	-165.6	1.6	3.3	-160.3

Table 1. Results of `ShapleyBO` for exemplary iterations 1 and 2 of BO with risk-averse confidence bound (Equation 4) on heteroscedastic target function $g(\boldsymbol{\theta})$ (Equation 7).

j	$\overline{\phi_j}(\hat{\mu})$	$\overline{\phi_j}(\hat{\sigma})$	$\overline{\phi_j}(\hat{\epsilon})$
1	-48.48	4.20	-87.27
2	-157.73	6.88	0.24

Table 2. Results of `ShapleyBO` averaged over all 60 iterations and all 30 BO restarts with risk-averse confidence bound (Equation 4) on heteroscedastic target function $g(\boldsymbol{\theta})$ (Equation 7).

uncertainty to prioritizing mean reduction, leading BO to favor configurations that perform well over those with high uncertainty. This transition is marked by a crossing in the contribution curves, see Figure 2, indicating a preference for mean reduction over uncertainty reduction.

Heteroscedastic Target Function: Secondly, we illustrate `ShapleyBO` on a two-dimensional ellipsoid function with noise depending on $\boldsymbol{\theta}$, see Section 4 for details. That is, we minimize $g(\boldsymbol{\theta}) + \boldsymbol{\epsilon}(\boldsymbol{\theta})$, where

$$\begin{aligned} g\colon\ [-15, 15]^2 &\to \mathbb{R}_0^+ \\ \boldsymbol{\theta} &\mapsto g(\boldsymbol{\theta}) = \sum_{i=1}^{2} i \cdot \theta_i^2 \end{aligned} \tag{6}$$

and

$$\boldsymbol{\epsilon}(\boldsymbol{\theta}) = 30 \cdot |\theta_1 - 15| + 0.3 \cdot |\theta_2 - 15|. \tag{7}$$

The noise grows strongly in θ_1, but only moderately in θ_2. Figure 3 shows contours of $g(\boldsymbol{\theta}) + \boldsymbol{\epsilon}(\boldsymbol{\theta})$. It becomes evident that the function varies stronger w.r.t. θ_2

than w.r.t. θ_1, while the noise is strongly affected by θ_1 and almost constant in θ_2. Hence, we expect the respective Shapley values for aleatoric (see Equation 4) uncertainty contributions to be high for θ_1 and low for θ_2, and vice versa for exploitation (mean optimization).

We run BO on $g(\boldsymbol{\theta}) + \boldsymbol{\epsilon}(\boldsymbol{\theta})$ with risk-averse confidence bound (*racb*), see Equation 4; we again average over 30 restarts of BO with 60 iterations each. `ShapleyBO` delivers contributions for each θ_j to each of *racb*'s components in each of BO's iterations. Table 1 has the results for exemplary iterations 1 and 2. Table 2 shows the contributions averaged over all $i \in \{1, \dots, 60\}$ iterations and all $r \in \{1, \dots, 30\}$ restarts. For instance, the averaged mean contributions of parameter j are

$$\overline{\phi_j}(\hat{\mu}) = \frac{1}{30} \sum_{r=1}^{30} \frac{1}{60} \sum_{i=1}^{60} \phi_{j,r,i}(\hat{\mu}). \tag{8}$$

It becomes evident that θ_2 is more important for the mean minimization than θ_1, while the latter contributes more to aleatoric uncertainty (noise) avoidance.

Summing up, the applications on both homo- and heteroscedastic target functions demonstrated that `ShapleyBO` manages to disentangle contributions of different parameters to different objectives of BO, thus providing valuable insights both into BO's inner working (see Figures 1, 2 and Table 1) and about the target function itself (see Table 2).

6 Shapley-Assisted Human Machine Interface

The ability to interpret BO can be particularly useful for Human-In-the-Loop (HIL) applications, where users observe each step in the sequential optimization procedure. In this case, `ShapleyBO` can inform users online; that is, while the optimization is still running, about why certain actions were chosen over others, instead of providing such explanations after the experiment has concluded. More specifically, we consider a human-AI collaborative framework [14,25,48,4,11], in which users can actively participate in the optimization by rejecting BO proposals and instead take actions on their own.

As demonstrated in Section 5, Shapley values can provide structural insights on the relative importance of parameters for the optimization by filtering out uncertainty contributions, see $\overline{\phi_j}(\hat{\mu})$ in Table 2 for instance. Our general hypothesis is that basing the decision to intervene on this information will speed up the optimization. The underlying idea is that users can reject proposals in case the respective Shapley values do not align with the user's knowledge about the optimization problem.

To test this hypothesis, we benchmark a `ShapleyBO`-assisted human-AI team against teams without access to Shapley values. To better illustrate this, we consider the real-world use case of personalizing control parameters of a wearable, assistive back exosuit by BO.

6.1 Personalizing Soft Exosuits

Wearable robotic devices, such as exoskeletons and exosuits, have emerged as promising tools in mitigating risk of injury and aiding rehabilitation [42,47]. With an increase in use cases and accessibility to a broader community, it has become apparent that the benefits of such devices can vary substantially between individuals. Besides design choices, which have to be made early on and are therefore often guided by (average) user anthropometrics, important factors influencing device efficacy are the magnitude and timing of assistance.

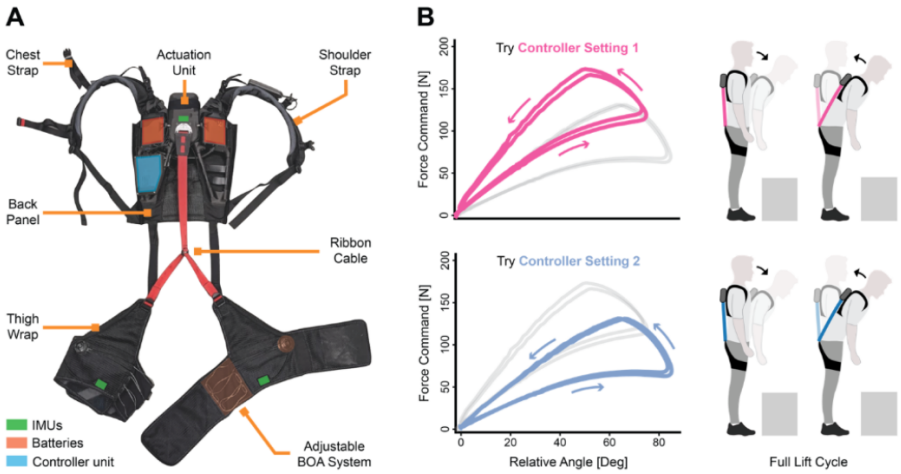

Fig. 4. A: Assistive soft back exosuit. **B**: Force profile example for preference learning. Subjects are asked to compare controllers setting 1 (pink) to 2 (blue). Each option varies in the amount of lowering gain (θ_{low}) and lifting gain (θ_{lif}), see [5].

To understand which settings work best for an individual, many studies follow HIL frameworks. These approaches comprise a feedback loop in which the impact of a controller modification on the objective function of interest is measured in real-time, and used to determine a set of control parameters that are likely to improve upon the current optimum in the subsequent iteration. Given that under such conditions there is typically no known analytical relationship between control inputs and objective function outputs, sample efficient, query based methods like BO have had considerable success for such applications [54,20].

6.2 Experimental Setup

Here, we explore the potential of `ShapleyBO` for the use-case of preference-based assistance optimization for a soft back exosuit, see Figure 4. To this end we consider a dataset in which 15 healthy individuals performed a simple, stoop

Algorithm 2 Human-AI Collaborative BO

create an initial design $D = \{(\boldsymbol{\theta}^{(i)}, \Psi^{(i)})\}_{i=1,\dots,n_{init}}$
while termination criterion is not fulfilled **do**
  train SM on data D
  propose $\boldsymbol{\theta}^{new}$ that optimizes $AF(SM(\boldsymbol{\theta}))$
  If intervention criterion is fulfilled
  $\boldsymbol{\theta}^{new} \leftarrow \boldsymbol{\theta}^{human}$
  End If
  evaluate Ψ on $\boldsymbol{\theta}^{new}$
  update $D \leftarrow D \cup (\boldsymbol{\theta}^{new}, \Psi(\boldsymbol{\theta}^{new}))$
end while
return $\arg\min_{\boldsymbol{\theta} \in D} \Psi(\boldsymbol{\theta})$

lifting task with a light (2kg) external load [5]. Details on the dataset can be found in the supplementary material. Preference was queried in a forced-choice paradigm. That is, within each iteration, participants were consecutively exposed to two control parameter settings and asked to indicate which of the two options they preferred for completing the given task. Each of the settings comprised two parameters, referred to as lowering gain θ_{low} and lifting gain θ_{lif}, which govern the amount of lowering and lifting assistance provided by the device, respectively, see also Figure 4.

This preference feedback was used to compute a posterior utility distribution over the considered parameter domains, relying on a probit likelihood model and a GP prior over the latent user utility as described in [16]. The experiment comprised three separate optimization blocks, in each of which the optimization was running for 12 iterations. To test `ShapleyBO` on this dataset, we averaged the three utility functions for each participant and interpolated by another GP to simulate the user's ground truth utility function. The detailed specifications for both the original BO and the auxiliary BO modeling the human can be found in the supplementary material and in our codebase.

The remaining setup in our experimental study closely follows the one in [48]. That is, the human can intervene in BO by rectifying proposals made by the algorithm, see pseudo code in Algorithm 2. We will compare our `ShapleyBO`-assisted human-BO team against the team in [48, Algorithm 1] and three other baselines (human alone, BO alone, human-BO team with different intervention criterion). We model human decisions by another BO, following [11,48]. This means that $\boldsymbol{\theta}^{human} = (\theta_{lif}^{human}, \theta_{low}^{human})^T$ is found by optimizing an AF modeling human preferences. We use a BO with the same SM and AF as for the outer loop, but with different exploration-exploitation preference and different initial design, representing differing risk-aversion and knowledge of the human, respectively.

All agents (A0, A1, A2, A3, A4) are equal to each other, the only difference being that the `ShapleyBO`-assisted agent intervenes based on Shapley values (A4), while the other agents intervene in each k-th iteration (A3) [48], based on the proposed parameters (A2), always (A1) or completely abstain from interven-

Agent	A0	A1	A2	A3	A4
	BO	Human	Param-Team	Venkatesh et al. [48]	Shap-Team
IC	never	always	$\theta_{lif}^{new}, \theta_{low}^{new}$	k-th iteration	$\phi_{lif}^{new}(\hat{\mu}), \phi_{low}^{new}(\hat{\mu})$

Table 3. Intervention Criteria (ICs) for `ShapleyBO`-assisted A4 and baselines A0-A3.

ing (A0), see overview in Table 3. A4 has access to `ShapleyBO` and bases their decision to intervene (line 5 in Algorithm 2) on the alignment of the Shapley values of a BO proposal $\boldsymbol{\theta}^{new} = (\theta_{lif}^{new}, \theta_{low}^{new})$ with the agent's knowledge. More precisely, A4 accepts a BO proposal θ (does not intervene) if

$$\frac{1}{\beta} < \frac{\phi_{lif}^{new}(\hat{\mu})}{\phi_{low}^{new}(\hat{\mu})} \Big/ \frac{1}{T}\sum_{t=1}^{T} \frac{\phi_{lif}^{human}(\hat{\mu})_t}{\phi_{low}^{human}(\hat{\mu})_t} < \beta, \tag{9}$$

where $t \in \{1, \ldots, T\}$ are iterations of the BO modeling the agent and $\phi(\hat{\mu})$ the Shapley mean contributions of $(\theta_{lif}, \theta_{low})$, i.e., the exosuit's lifting and lowering gain, respectively. We discuss different Shapley-based intervention criteria in the supplement. For A2's intervention criterion we consider the alignment of $(\theta_{lif}^{new}, \theta_{low}^{new})$ with the agents knowledge based on the parameter values itself. That is, A3 accepts a BO proposal (does not intervene) if

$$\frac{1}{\beta} < \frac{\theta_{lif}^{new}}{\theta_{low}^{new}} \Big/ \frac{1}{T}\sum_{t=1}^{T} \frac{\theta_{lif,t}^{human}}{\theta_{low,t}^{human}} < \beta. \tag{10}$$

6.3 Results

We simulate 40 personalization rounds with 10 iterations and initial design of size 3 each for all five agents. We compare them with respect to *optimization paths*, which show best incumbent target values (utility) in a given iteration. The BO uses GP as AM and cb with $\lambda = 20$ as AF; the BO modelling human proposals uses GP and cb with $\lambda = 200$ and prior knowledge of 90 data points. For further details on the experiments, please refer to the supplementary material.

Figure 5 exemplarily summarizes results for individual 1; results for remaining 14 individuals as well as experimental details can be found in the supplement. For all 15 subjects, `ShapleyBO`-assisted A4 (Shap-team) on average outperforms human and BO baseline as well as [48] and a team that bases their decision to intervene on the proposed parameters. This latter comparison particularly confirms that Shapley values are a meaningful measure for human-BO alignment that cannot be replaced by another notion of alignment without loss of efficiency. For 10 (among whom is subject 1, see Figure 5) out of 15 subjects the observed outperformance of the `ShapleyBO`-assisted A4 over competitors is significant at 95% confidence level.

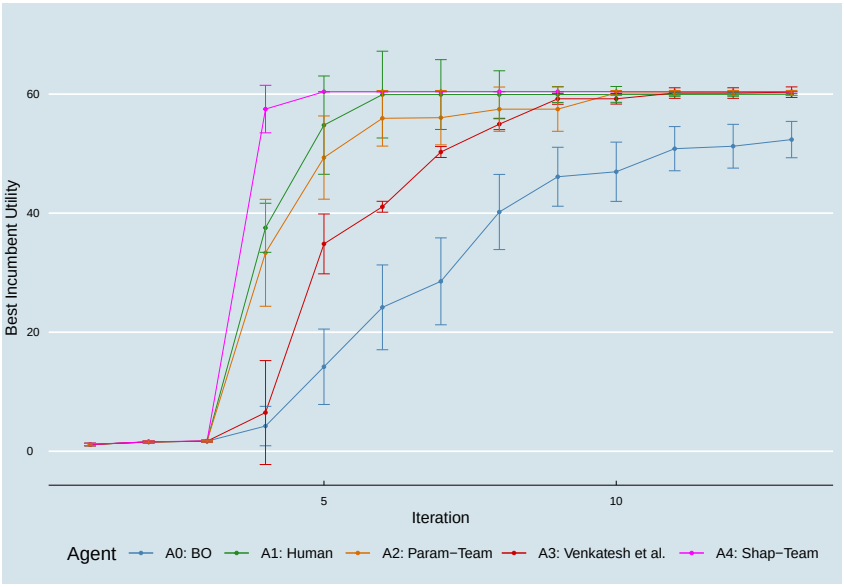

Fig. 5. Results of Agents A0-A4 (see Table 3) in human-AI collaborative BO for simulated exosuit personalization (individual 1) with 10 iterations and 3 initial samples each. Error bars indicate 95% confidence intervals; $k = 2$ for A3, $\beta = 2$ for A2 and A4. Results for remaining individuals can be found in the supplementary material.

7 Discussion

By quantifying the contribution of each parameter to the proposals, `ShapleyBO` aids in the communication of the rationale behind specific optimization decisions. This interpretability is not only crucial for trust in HIL applications, it also enhances their efficiency in a human-AI collaborative setup. The use case of customizing exosuits illustrates the practical benefits of this approach, suggesting that ShapleyBO is a valuable practical tool for personalizing soft back exosuits.

More generally and beyond exosuits, we conclude that `ShapleyBO` fosters more efficient human-AI collaboration by serving as an explanation interface between the optimization algorithm and humans.

Our paper opens up a multitude of directions for future work. The simulation results in Section 6 based on real-world data motivate the actual deployment of Shapley-assisted human-in-the-loop optimization of exosuits in a user study. To this end, the intervention logic used in the simulation study could benefit from an intuitive or visual explanation interface.

On the methodological end, extensions of `ShapleyBO` to multi-criteria BO appear straightforward, as long as additive AFs are used. Moreover, the stability of the Shapley attributions under different BO settings (e.g., kernel or mean choice in Gaussian process, noise, random seeds) might be investigated. A sensitivity analysis along the lines of [36,37,38,39,40] could yield fruitful insights into the

robustness of `ShapleyBO`. Moreover, a direct integration of preferences similar to CoExBo [3] might increase the efficiency of human-machine interaction further in a collaborative BO setup.

What is more, a thorough mathematical study of the sublinear regret bounds of BO with confidence bound [44] under Shapley-assisted human interventions might foster theoretical understanding of why Shapley-assisted teams outperform competitors. The theoretical results on general human-AI collaborative BO [48], technically relying on Sobolev spaces, can serve as a starting point for an extended study that explicitly accounts for the `ShapleyBO`-assisted human-machine interface.

Acknowledgments. We thank all three anonymous reviewers for valuable feedback on our work. TA and JR gratefully acknowledge support by the Federal Statistical Office of Germany within the co-operation project "Machine Learning in Official Statistics". JR further acknowledges support by the Bavarian Academy of Sciences (BAS) through the Bavarian Institute for Digital Transformation (bidt) and by the LMU mentoring program of the Faculty of Mathematics, Informatics, and Statistics. YS is supported by the DAAD program Konrad Zuse Schools of Excellence in Artificial Intelligence, sponsored by the Federal Ministry of Education and Research. EH has received funding from the European Union's Horizon Europe research and innovation programme under the Marie Sklodowska-Curie grant agreement No 101073307.

Disclosure of Interests. CJW is an inventor of at least one patent application describing the exosuit components described in the paper that have been filed with the U.S. Patent Office by Harvard University. Harvard University has entered into a licensing agreement with Verve Inc., in which CJW has an equity interest and a board position. The other authors declare that they have no competing interests.

Ehtical Statement The data used to estimate the utility functions was collected as part of another study [5] for which ethical approval was obtained from the applicable institutional ethical review board.

References

1. Aas, K., Jullum, M., Løland, A.: Explaining individual predictions when features are dependent: More accurate approximations to shapley values. Artificial Intelligence **298**, 103502 (2021)
2. Adachi, M., Chau, S.L., Xu, W., Singh, A., Osborne, M.A., Muandet, K.: Bayesian optimization for building social-influence-free consensus. arXiv preprint arXiv:2502.07166, accessed 29 May 2025 (2025)
3. Adachi, M., Planden, B., Howey, D.A., Maundet, K., Osborne, M.A., Chau, S.L.: Looping in the human: Collaborative and explainable bayesian optimization. In: 27th International Conference on Artificial Intelligence and Statistics (AISTATS). vol. 238. PMLR (2024)
4. Akata, Z.e.a.: A research agenda for hybrid intelligence: Augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence. Computer **53**(8), 18–28 (2020)

5. Arens, P., Quirk, D.A., Pan, W., Yacoby, Y., Doshi-Velez, F., Walsh, C.J.: Preference-based assistance optimization for lifting and lowering with a soft back exosuit. Science Advances **11**(15) (2025)
6. Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al.: Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. Information fusion **58**, 82–115 (2020)
7. Arun Kumar, A.V., Shilton, A., Gupta, S., Rana, S., Greenhill, S., Venkatesh, S.: Enhanced bayesian optimization via preferential modeling of abstract properties. arXiv preprint arXiv:2402.17343, accessed May 28 2025 (2024)
8. Arun Kumar, A.V., Shilton, A., Gupta, S., Ryan, S., Abdolshah, M., Le, H., Rana, S., Berk, J., Rashid, M., Venkatesh, S.: Accelerated experimental design using a human–ai teaming framework. Knowledge-Based Systems **315**, 113138 (2025)
9. Bakshi, S., Sharma, S., Khanna, R.: Shapley-value-based hybrid metaheuristic multi-objective optimization for energy efficiency in an energy-harvesting cognitive radio network. Mathematics **11**(7), 1656 (2023)
10. Bennetot, A., Donadello, I., El Qadi El Haouari, A., Dragoni, M., Frossard, T., Wagner, B., Sarranti, A., Tulli, S., Trocan, M., Chatila, R., et al.: A practical tutorial on explainable ai techniques. ACM Computing Surveys **57**(2), 1–44 (2024)
11. Borji, A., Itti, L.: Bayesian optimization explains human active search. Advances in Neural Information Processing Systems **26** (2013)
12. Breiman, L.: Random forests. Machine Learning **45**(1), 5–32 (2001)
13. Burton, J.W., Stein, M.K., Jensen, T.B.: Beyond algorithm aversion in human-machine decision-making. In: Judgment in Predictive Analytics, pp. 3–26. Springer (2023)
14. Chakraborty, T., Seifert, C., Wirth, C.: Explainable bayesian optimization. arXiv preprint arXiv:2401.13334, accessed May 29 2025 (2024)
15. Chau, S. L., Muandet, K., Sejdinovic, D.: Explaining the Uncertain: Stochastic Shapley Values for Gaussian Process Models. In: Advances in Neural Information Processing Systems, 36, 50769-50795. (2023)
16. Chu, W., Ghahramani, Z.: Preference learning with gaussian processes. In: Proceedings of the 22nd International Conference on Machine Learning. pp. 137–144 (2005)
17. Cisse, A., Evangelopoulos, X., Carruthers, S., Gusev, V.V., Cooper, A.I.: Hypbo: Expert-guided chemist-in-the-loop bayesian search for new materials. arXiv preprint arXiv:2308.11787, accessed 10 Jun 2025 (2023)
18. Croppi, F.: Explaining sequential model-based optimization (2021), master thesis, LMU Munich
19. Dietvorst, B.J., Simmons, J.P., Massey, C.: Algorithm aversion: people erroneously avoid algorithms after seeing them err. Journal of Experimental Psychology: General **144**(1), 114 (2015)
20. Ding, Y., Kim, M., Kuindersma, S., Walsh, C.J.: Human-in-the-loop optimization of hip assistance with a soft exosuit during walking. Science Robotics **3**(15), eaar5438 (2018)
21. Fisher, A., Rudin, C., Dominici, F.: All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. Journal of Machine Learning Research **20**(177), 1–81 (2019)
22. Frazier, P., Wang, J.: Bayesian optimization for materials design. In: Information Science for Materials Discovery and Design, pp. 45–75. Springer (2016)

23. Fréchette, A., Kotthoff, L., Michalak, T., Rahwan, T., Hoos, H., Leyton-Brown, K.: Using the shapley value to analyze algorithm portfolios. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 30 (2016)
24. Friedman, J.H.: Greedy function approximation: A gradient boosting machine. The Annals of Statistics **29**(5), 1189–1232 (2001)
25. Gupta, S., Shilton, A., AV, A.K., Ryan, S., Abdolshah, M., Le, H., Rana, S., Berk, J., Rashid, M., Venkatesh, S.: Bo-muse: A human expert and ai teaming framework for accelerated experimental design. arXiv preprint arXiv:2303.01684, accessed 10 Jun 2025 (2023)
26. Hüllermeier, E., Waegeman, W.: Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. Machine Learning **110**(3), 457–506 (2021)
27. Jansen, C., Schollmeyer, G., Blocher, H., Rodemann, J., Augustin, T.: Robust statistical comparison of random variables with locally varying scale of measurement. In: Uncertainty in Artificial Intelligence (UAI). Proceedings of Machine Learning Research (PMLR). pp. 941–952 (2023)
28. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Advances in neural information processing systems (NeurIPS), Curran Associates, Inc. (2017)
29. Makarova, A., Usmanova, I., Bogunovic, I., Krause, A.: Risk-averse heteroscedastic bayesian optimization. In: Advances in Neural Information Processing Systems **34**, 17235–17245 (2021)
30. Močkus, J.: On Bayesian methods for seeking the extremum. In: Optimization techniques IFIP technical conference. pp. 400–404. Springer (1975)
31. Paleja, R., Ghuy, M., Ranawaka Arachchige, N., Jensen, R., Gombolay, M.: The utility of explainable ai in ad hoc human-machine teaming. In: Advances in Neural Information Processing Systems. vol. 34, pp. 610–623. Curran Associates, Inc. (2021)
32. Peters, H.: Game theory: A Multi-leveled approach. Springer (2015)
33. Psaros, A.F., Meng, X., Zou, Z., Guo, L., Karniadakis, G.E.: Uncertainty quantification in scientific machine learning: Methods, metrics, and comparisons. Journal of Computational Physics **477**, 111902 (2023)
34. Pyzer-Knapp, E.O.: Bayesian optimization for accelerated drug discovery. IBM Journal of Research and Development **62**(6) (2018)
35. Rafailov, R., Sharma, A., Mitchell, E., Manning, C.D., Ermon, S., Finn, C.: Direct preference optimization: Your language model is secretly a reward model. Advances in Neural Information Processing Systems **36** (2024)
36. Rodemann, J.: Robust Generalizations of Stochastic Derivative-Free Optimization. Master's thesis, LMU Munich (2021)
37. Rodemann, J., Augustin, T.: Accounting for imprecision of model specification in bayesian optimization. Poster presented at International Symposium on Imprecise Probabilities (ISIPTA) (2021)
38. Rodemann, J., Augustin, T.: Accounting for gaussian process imprecision in bayesian optimization. In: International Symposium on Integrated Uncertainty in Knowledge Modelling and Decision Making. pp. 92–104. Springer (2022)
39. Rodemann, J., Augustin, T.: Imprecise bayesian optimization. Knowledge-Based Systems **300**, 112186 (2024)
40. Rodemann, J., Fischer, S., Schneider, L., Nalenz, M., Augustin, T.:Not all data are created equal: Lessons from sampling theory for adaptive machine learning. In: International Conference on Statistics and Data Science (ICSDS) (2022)

41. Shapley, L.S.: A value for n-person games. Contributions to the Theory of Games **2**(28), 307–317 (1953)
42. Siviy, C., Baker, L.M., Quinlivan, B.T., Porciuncula, F., Swaminathan, K., Awad, L.N., Walsh, C.J.: Opportunities and challenges in the development of exoskeletons for locomotor assistance. Nature Biomedical Engineering **7**(4), 456–472 (2023)
43. Snoek, J., Larochelle, H., Adams, R.P.: Practical Bayesian optimization of machine learning algorithms. In: Advances in Neural Information Processing Systems **25**, 2951–2959 (2012)
44. Srinivas, N., Krause, A., Kakade, S.M., Seeger, M.W.: Information-theoretic regret bounds for gaussian process optimization in the bandit setting. IEEE transactions on Information Theory **58**(5), 3250–3265 (2012)
45. Štrumbelj, E., Kononenko, I.: Explaining prediction models and individual predictions with feature contributions. Knowledge and Information Systems **41**(3), 647–665 (2014)
46. Tallón-Ballesteros, A., Chen, C.: Explainable ai: Using shapley value to explain complex anomaly detection ml-based systems. Machine Learning and Artificial Intelligence **332**, 152 (2020)
47. Toxiri, S., Näf, M.B., Lazzaroni, M., Fernández, J., Sposito, M., Poliero, T., Monica, L., Anastasi, S., Caldwell, D.G., Ortiz, J.: Back-support exoskeletons for occupational use: an overview of technological advances and trends. IISE Transactions on Occupational Ergonomics and Human Factors **7**(3-4), 237–249 (2019)
48. Venkatesh, A.K., Rana, S., Shilton, A., Venkatesh, S.: Human-ai collaborative Bayesian optimisation. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems. vol. 35, pp. 16233–16245 (2022)
49. Wang, Y.C., Chen, T.: Adapted techniques of explainable artificial intelligence for explaining genetic algorithms on the example of job scheduling. Expert Systems with Applications **237**, 121369 (2024)
50. Watson, D.S., O'Hara, J., Tax, N., Mudd, R., Guy, I.: Explaining predictive uncertainty with information theoretic shapley values. arXiv preprint arXiv:2306.05724, accessed May 3 2025 (2023)
51. Wu, H.C.: Solving fuzzy optimization problems using shapley values and evolutionary algorithms. Mathematics **11**(24), 4871 (2023)
52. Xiao, H., Wang, Z., Zhu, Z., Zhou, J., Lu, J.: Shapley-nas: Discovering operation contribution for neural architecture search. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11892–11901 (2022)
53. Xu, W., Adachi, M., Jones, C.N., Osborne, M.A.: Principled bayesian optimization in collaboration with human experts. In: Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., Zhang, C. (eds.): Advances in Neural Information Processing Systems (NeuIPS). vol. 37, pp. 104091–104137. Curran Associates, Inc. (2024)
54. Zhang, J., Fiers, P., Witte, K.A., Jackson, R.W., Poggensee, K.L., Atkeson, C.G., Collins, S.H.: Human-in-the-loop optimization of exoskeleton assistance during walking. Science **356**(6344), 1280–1284 (2017)