


Bayesian uncertainty quantification for machine-learned models in physics

Yarin Gal, Petros Koumoutsakos , Francois Lanusse, Gilles Louppe and Costas Papadimitriou

Abstract | Being able to quantify uncertainty when comparing a theoretical or computational model to observations is critical to conducting a sound scientific investigation. With the rise of data-driven modelling, understanding various sources of uncertainty and developing methods to estimate them has gained renewed attention. Five researchers discuss uncertainty quantification in machine-learned models with an emphasis on issues relevant to physics problems.

“Every time a scientific paper presents a bit of data, it’s accompanied by an error bar — a quiet but insistent reminder that no knowledge is complete or perfect,” wrote astrophysicist Carl Sagan in the 1995 book *The Demon-Haunted World: Science as a Candle in the Dark*. Sagan’s words resonate even more today as physicists are increasingly relying on machine-learned models. These use complex statistical methods and large amounts of training data to make predictions without having a pre-specified model to do so. The models are powerful, but quantifying their uncertainties is challenging.

 *What is uncertainty estimation and why is it important in physics?*

Petros Koumoutsakos, Costas Papadimitriou

Physics strives to acquire knowledge about the world through a quest for quantifiable relationships between observations and ideas. These relationships are often expressed with mathematical and computational models describing physical principles such as conservation laws. In recent years, models learned from data have gained substantial attention. Unprecedented computational power has made both computation and data modelling approaches a necessity for prediction and decision making across science and technology. However, limited or information-poor data and unknown physical phenomena that affect a system, but are not captured by its models,


mean that all predictions are uncertain and there is risk associated with every decision. The estimation of uncertainty is a unifying theme and a fundamental aspect of modelling.

A prominent method for linking models and observations and estimating uncertainties is Bayesian inference. As an inductive method for learning from finite data, it is intimately linked to the field of machine learning. It is distinguished by the use of an a priori degree of belief that is assigned to the model parameters, expressed in terms of prior probability distributions. The classical Bayes formula updates these probabilities systematically in the light of new data, and in that context, it is closely related to the intuitive reasoning of physicists.

Francois Lanusse Uncertainty quantification underpins most of modern fundamental physics, which at its core seeks to compare theoretical models to observations. Although approaches may vary depending on the fields (for example, frequentist statistics in particle physics, Bayesian statistics in cosmology), statistical uncertainty quantification methods always aim to make the comparison of models to observations quantitative, in order to prefer or reject particular theories. In this sense, robust and reliable uncertainty quantification is of paramount importance to physicists, who will often prefer an experiment or methodology with less constraining power if it can lead to a better control of the associated uncertainties and

more trustworthy results. This importance placed in the robustness and trustworthiness of statistical methods has no small role in the relative scepticism that modern machine learning has been facing in many areas of physics. Neural networks are often still perceived as uninterpretable black boxes with dubious uncertainties, which brings reluctance to use them even if they appear to outperform more classical analysis techniques. In fact, neural networks can be interpreted as sound probabilistic models, with asymptotic convergence guarantees (for example, infinite data, infinite network size). In that, they are no different than more traditional inference techniques, such as Markov chain Monte Carlo (MCMC), which are also only asymptotically correct and can in practice be plagued by a large number of technical issues, but are nonetheless considered the gold standard in fields such as cosmology.

Building a deep understanding of the meaning of neural networks in a probabilistic context, as well as experience with using these models in the non-asymptotic regime, is slowly gathering pace in different areas of the physical sciences, in which it is more critical than in many other common application domains of deep learning. One of the most compelling examples of using neural networks under a sound statistical framework is the recent growth of simulation-based inference techniques¹ based on neural density estimation.

 *What are the tools for uncertainty estimation in machine learning and deep learning?*

P.K., C.P. Bayesian uncertainty quantification is distinguished by its high computational cost due to the need to represent the posterior uncertainty in a multidimensional parameters space and to evaluate multidimensional integrals over the model parameters, to estimate: first, the posterior probability of quantities of interest that are important for decision making; second, the relative plausibility of different physical models, by computing and comparing their evidence; and third, the systematic incorporation of heterogeneous data (such as different physical properties) through hierarchical

The contributors

Yarin Gal is Associate Professor of Machine Learning at the University of Oxford Computer Science department, UK, and leads the Oxford Applied and Theoretical Machine Learning (OATML) group. He has made substantial contributions to early work in modern Bayesian deep learning — quantifying uncertainty in deep learning — and developed machine learning and artificial intelligence tools that can inform their users when the tools are ‘guessing at random’. These tools have been deployed widely in industry and academia, such as in medical applications, robotics, computer vision, astronomy and sciences, and are also used by NASA.

Petros Koumoutsakos is the Herbert S. Winokur Professor of Science and Engineering in the John A. Paulson School of Engineering and Applied Sciences at Harvard University. He is elected Fellow of the American Society of Mechanical Engineers (ASME), the American Physical Society (APS), the Society of Industrial and Applied Mathematics (SIAM) and a recipient of the SIAM/ACM Gordon Bell award in Supercomputing. He is an international member of the US National Academy of Engineering (NAE). Petros’ research interests are on the fundamentals and applications of computing and artificial intelligence to understand, predict and optimize complex systems in engineering, nanotechnology and medicine.

Francois Lanusse is a CNRS researcher, part of CosmoStat Laboratory at CEA Saclay. Previously, he worked at the Berkeley Center for Cosmological Physics, the Foundation of Data Analysis Institute at UC Berkeley and in the McWilliams Center for Cosmology at Carnegie Mellon University. His research is at the intersection of cosmology and machine learning.

Gilles Louppe is Associate Professor in Artificial Intelligence and Deep Learning at the University of Liège in Belgium. His research is at the intersection of deep learning, approximate inference and the physical sciences. He has been developing a new generation of simulation-based inference algorithms based on deep learning, with several applications in particle physics, astrophysics, astronomy and gravitational wave science.

Costas Papadimitriou is Professor of Structural Dynamics and the founding Director of System Dynamics Laboratory at the University of Thessaly in Greece. His research interests include data-driven uncertainty quantification in engineering and applied sciences, optimal experimental design, structural health monitoring and reliability. He is the recipient of the 2014 European Association of Structural Dynamics (EASD) Senior Award in Computational Structural Dynamics.

Bayesian inference. The model parameters can be kept small by incorporating physical principles and symmetries or by identifying low-dimensional manifolds for the evolution of the quantities of interest that are evaluated by these models.

Over the years, several computational methods such as MCMC methodologies, well-known to physicists, have been developed to sample effectively the posterior distributions or estimate the corresponding integrals. Alternatively, Laplace approximation techniques provide estimates based on maximum a posteriori estimates and local representations of uncertainty. These tools have been originally developed for Bayesian uncertainty quantification in classical statistical models and are now being adapted to machine-learned models. Examples include adaptations of variational inference and dropout MCMC (see REFS. 2,3 for comprehensive reviews).

Yarin Gal The field of Bayesian deep learning has experienced a boom in research in the past few years, with various tools developed to estimate different types of uncertainty, each with its own properties. These tools can be broadly categorized

following different axes, and understanding these is critical to selecting the right tool for an application. Tools can be divided in the following categories: first, the type of uncertainty they capture (epistemic or aleatoric, see BOXES 1 and 2); second, by the computational constraints and requirements (such as many forward passes versus a single one, or memory requirements to store many models versus a single model); third, by the ease of use (how much ‘specialist knowledge’ in machine learning is needed to use the tool versus being ‘off the shelf’); fourth, by the justification underlying them (principled versus ad hoc, grounded in statistics or not); fifth, by the statistical paradigm (Bayesian versus frequentist); and sixth, if approximating a Bayesian posterior, by the type of approximation and where they lie on the approximation spectrum (for example, variational versus MCMC approximation, ‘crude and cheap’ versus ‘expressive and expensive’ approximation), and others.

Take the example of ‘dropout as variational inference’⁴ in Bayesian neural networks. In terms of the above categories, the method can capture both epistemic and aleatoric uncertainty (first category), requires multiple forward passes and stores only a single model in memory (second

category), can be used ‘off the shelf’ for models with dropout layers by simply doing several forward passes with dropout turned on at test time (third category), follows a Bayesian paradigm with a principled variational approximation if the dropout parameter value is set appropriately (fourth and fifth categories), and can be seen as a ‘cheap and crude’ approximation, which may be sufficient for some applications, but may be insufficient for others (sixth category).

Compare this to the ‘deep ensemble uncertainty’ method, which requires storing multiple models in memory (second category), can be used ‘off the shelf’ if one can train additional models (third category), follows ad hoc intuition rather than statistical justification (fourth category, although some work has attempted to ground it under additional assumptions), and can be seen as a slightly more ‘expressive and expensive’ approximation, which may improve performance for some applications if one has access to additional computational resources (sixth category).

Two more representative examples along these axes are Hamiltonian Monte Carlo for inference in Bayesian neural networks⁵ and deterministic uncertainty quantification (DUQ)⁶: the former is the most computationally expensive of the methods mentioned above, whereas the latter is the cheapest. Hamiltonian Monte Carlo requires multiple forward and backward passes while storing many model replicas in memory (second category), requires specialist knowledge in machine learning to use effectively (third category), and follows a Bayesian paradigm with a rigorous MCMC approximation (fourth and fifth categories). By contrast, DUQ requires only a single forward pass with a single model (second category), minimal changes to the training procedure (third category), but captures only epistemic uncertainty (first category) and follows ad hoc justification (more recent extensions of DUQ^{7,8} remedy these last points and ground the method as a principled approximation to a Gaussian process, fourth and fifth categories).

F.L. The tools needed to model aleatoric and epistemic uncertainties (see BOXES 1 and 2) are different. Aleatoric uncertainties are independent of the machine learning method. For instance, one can never exactly recover the true value of a quantity that is only observed through a few noisy measurements, no matter what methodology is used. Modelling these uncertainties essentially means modelling distributions,

for instance, the probability distribution of this unknown quantity given the observed data. This is an area in which deep learning has made substantial advances in recent years, under the general notions of neural density estimators and deep generative models (including generative adversarial networks, variational autoencoders, normalizing flows, diffusion models and autoregressive models). Whereas modelling high-dimensional distributions was once a problem plagued by the curse of dimensionality, recent state-of-the-art models are able to represent high-dimensional distributions over complex data, with examples ranging from the usual images of human faces found in the machine learning literature, to maps of the large-scale structure of the Universe, useful for cosmology. This ability to efficiently model arbitrary distributions with neural networks has in particular led to a recent development of so-called simulation-based inference techniques¹, which offer an alternative to standard MCMC techniques in parameter inference problems in which the physical model is only provided in the form of a numerical simulator.

Epistemic uncertainties are much more subtle to model in a physically meaningful way. These would refer, in this context, to the uncertainty on the neural network itself, and can be summarized by the following question: given a finite amount of training data, how reliable is the prediction made by a specific neural network model? This is the question that Bayesian neural network

techniques try to address. Note, however, that this estimation of epistemic uncertainties in neural networks is typically made under ad hoc priors on the network architecture and on the weights of the model, which are not directly interpretable in terms of a physically meaningful prior on the functional space of the neural network output. The resulting uncertainties should therefore be handled with care in a physical inference context, but they can still be useful to detect whether a given model is poorly constrained by data, in which case, more data can be acquired, as in active sampling schemes.

Q *What do we need to worry about when doing uncertainty estimation?*

Y.G. After understanding the constraints of the application (and thus which uncertainty tools are appropriate for the task), choosing metrics to quantify how well uncertainty is modelled is the next most critical point. Aleatoric uncertainty can be quantified using frequentist statistical tools such as expected calibration error, but epistemic uncertainty (an inherently subjective quantity) cannot. In fact, a model can be perfectly calibrated yet give meaningless epistemic uncertainty, and vice versa. Metrics for epistemic uncertainty quality include, for example, selective classification in which the accuracy of the model is evaluated only on the predictions with the lowest epistemic uncertainty (and predictions with high epistemic uncertainty are, for example, referred to a human to label).

F.L. In fundamental physics, and in particular in cosmology, uncertainties classically fall into two categories, statistical and systematic uncertainties, which broadly map to the concepts of aleatoric and epistemic uncertainties. Systematic uncertainties are the most worrisome as they can lead to invalid conclusions if they are not properly controlled and kept substantially smaller than the statistical uncertainties. In a physics analysis, the control of systematic uncertainties goes far beyond what is typically considered in the machine learning literature on epistemic uncertainties. The uncertainty on the fit of a neural network is only one link in a very long analysis chain, which covers, in particular, any errors or biases in the training set used to train the network, or unexpected contamination of the observational data on which the network is applied. These considerations are more distilled and apparent in physics, but are equally important (albeit sometimes overlooked) in any applications of deep learning.

P.K., C.P. Bayesian inference requires sampling of distributions with dimensionality greater than or equal to the number of model parameters. The sampling necessitates numerous evaluations, making the process computationally demanding, particularly when the underlying model itself is computationally expensive. Special care is necessary to develop sampling algorithms that harness the capabilities of modern supercomputers⁹. Moreover, the accurate processing of information available from heterogeneous sources of data, which may reflect both randomness in the signal and noise in the sensors, is a major challenge that needs to be addressed. Today, machine-learned models are not readily extended to account for data heterogeneity and correlations. Finally, for Bayesian uncertainty quantification, although priors for models derived from first principles may easily encode prior knowledge, such priors are not easy to develop for machine-learned models.

Gilles Louppe Before uncertainty quantification, the first step for a principled Bayesian analysis is to make sure the priors and the observational models together form a data-generating process that reflects domain knowledge adequately. (Note that here the observation models should not be mistaken with the ‘neural network models’. The former refers to the forward physical model that we are interested in and want to

Box 1 | Types of uncertainties

There are two types of uncertainty important for both physicists and machine learning scientists.

- Aleatoric uncertainty refers to the intrinsic uncertainty of a particular system and the observed data. It arises due to the intrinsic and irreducible stochastic variability in the data-generating process. Aleatoric uncertainty — or data uncertainty — cannot be readily reduced as it is inherent to the measurement data.
- Epistemic uncertainty refers to imperfections of the observational models used to describe physical phenomena. Epistemic uncertainty — or model uncertainty — arises from our ignorance about the underlying physical process itself reflecting our lack of knowledge about its structure or its parameters. In machine learning, epistemic uncertainty is associated with model structure, cost function and training algorithms, and can be reduced as more data become available. Optimal design of experiments can assist the allocation of sensing and data acquisition. Note that the ‘model’ in machine learning is not the same as the observational model.

Although distinguishing between the two types of uncertainties may help to understand the problem, perhaps the more important issue for decision making is the intended use of each type of uncertainty¹¹. Such uses and decisions often imply additional factors that go beyond data and model relationships that may be formalized by machine learning models.

In the physical sciences, systematic uncertainties also have an important role.

- Systematic uncertainty corresponds to known unknowns that affect the outcome. Contrary to data uncertainty, systematic uncertainties are usually not conceived as random fluctuations, but rather as static, yet unobserved, variables. In this sense, systematic uncertainties are a source of observational model uncertainty.

Box 2 | An intuitive example

Gilles Louppe To illustrate the uncertainties discussed in BOX 1, imagine a pitcher throwing a baseball. The observational model in this case is given by Newton's laws of motion ($F = ma$), which describe the trajectory of the ball. The parameters of the model are the initial angle (θ) and the initial velocity (v). The final position where the ball lands can only be measured approximately (for example, with a precision of 10 cm). For a fixed initial angle and velocity, the model can generate multiple observations. It is first deterministic, but then can be made stochastic to capture the imprecision of the measurements.

In this setting, the aleatoric uncertainty is the uncertainty due to the imperfect measurements. It would not be reduced by having more data, but it could be reduced if one could make better measurements. For the (parametric) epistemic uncertainty, if the initial angle and the initial velocity are unknown, then in a Bayesian framework one can place a prior on the parameter values. This prior should reflect prior knowledge (for instance, the initial velocity cannot be very high because the pitcher has a limit and the angle should point forward around 30–45°). Given one (or several) observations, Bayesian inference can be used to reduce the epistemic uncertainty about the model parameters.

Systematic uncertainties could also be accounted for to capture known unknowns through nuisance parameters. For example, air resistance or wind can easily be incorporated in the model, but their exact parametric values will be left unmeasured. This lack is not an obstacle to the analysis, however, as one can either marginalize them out (when we are Bayesians and place a prior over their values) or profiled out (when we are frequentist and assume the worst-case scenario).

Finally, the model is clearly wrong and therefore misspecified because there are plenty of factors not taken into account (the ball is approximated as a point, but it is in fact a deformable, rotating body; the air is not homogeneous and so on). Furthermore, if the initial velocity is close to the speed of light, the model will produce wrong predictions. Yet, this model remains useful because it still shows predictive performance in the regime that one is interested in. In that sense, it at least captures the relevant part of the true data-generating process for the analysis that one wants to make.

use during inference; the latter is typically used to invert the forward-generating process.) The observational model should capture the pertinent structure of the true data-generating process, whereas the prior model should be chosen to produce plausible outcomes, and can be diagnosed with prior predictive checks. Vague and uninformative prior models should be avoided if they lead, when combined with the observational model, to unrealistic outcomes that are inconsistent with the expertise of the domain.

Once the model is set, inference can be carried out in various ways. For most models, exact inference is not an option and one must rely on approximate (Bayesian) inference engines based on MCMC methods or on simulation-based inference. For the inferences to be reliable and meaningful, one must make sure that the results are computationally faithful (for example, using coverage diagnostics to make sure the posteriors are neither too conservative nor too overconfident). If not used properly, inference engines can indeed produce results that are quite far off from the ground truth posterior that one aims to estimate, which may have detrimental consequences. For example, in the physical sciences, in which the goal is often to constrain parameters of interest, wrongly excluding plausible values could drive the scientific enquiry

in the wrong direction. For these reasons, uncertainty quantification should always come with diagnostics designed to probe the correct behaviour of the inference method, such as \hat{R} diagnostics for MCMC or coverage diagnostics in simulation-based inference.

Finally, if a model is a good fit, then one should be able to use it to generate data that resemble the observed data. If posterior predictive checks reveal that the observed data are very unlikely, then it is certainly a sign of model misspecification — the model is wrong and fails to reproduce the pertinent structure of the true data-generating process. Accordingly, model criticism from inference results should be used to inform the next revision of the model. For example, this can be done by incorporating nuisance parameters to account for systematic uncertainties.

F.L. Physicists should be very careful when assigning a meaning to epistemic uncertainties provided by models such as Bayesian neural networks. Although these models are indeed Bayesian, a Bayesian posterior is only meaningful if the corresponding prior is itself meaningful, which is generally not the case when imposing priors on neural network weights. A safer approach is to use high epistemic uncertainties as a sufficient but not

necessary condition, to detect a poorly constrained model, typically due to lack of training data. This condition can be used to decide where to sample additional training data.

However, the most worrisome failure modes would come from unknown unknowns. A first class of examples would be anomaly or out-of-distribution detection, which remains a very challenging task for machine learning methods when the data are high dimensional. Thus machine learning models may not always be guaranteed to recognize extremely scientifically interesting yet rare and unexpected events as being new. A second class of examples would be covariate or distributional shifts, in which for instance, the observational data on which the network is applied can be contaminated by subtle and unexpected effects that are not present in the training data, and may not be detectable by standard techniques. Because the training data are no longer representative, the response of the neural network model may be biased. In particular, standard cosmological analyses have developed procedures and null tests to detect such contaminations, but these are based on an understanding of how the data are classically analysed (in terms of two-point correlation functions), which does not transfer directly to deep neural networks.

Q Will the integration of physics knowledge into the models help improve the level of uncertainty?

Y.G. Integrating physics knowledge in a model in the form of invariances — be it translational or rotational equivariance, the conservation of energy or hybrid models integrated with simulators — affects the predictions of the model, and its uncertainty. For example, a model that does not respect translational invariance, when trained, for example, for object detection tasks with objects appearing only in the bottom half of an image, would have high uncertainty given new examples with the same object appearing in the top half of the image. However, a model with translation invariance, given the same new example, would produce the same output that it would produce for the training examples where the object appeared at the bottom half of the image, and do so confidently. The invariances we choose to build into our models correspond to our assumptions about what constitutes 'new examples' that should have high uncertainty.

G.L. Physics knowledge is helpful for inference engines. For example, in simulation-based inference with deep neural networks, inductive biases from physics knowledge can be used to substantially reduce the number of simulations necessary to produce accurate results. Depending on the forward process and on the power of the inductive bias, the gains in efficiency and accuracy can be of several orders of magnitude.

P.K., C.P. The exclusive development of models derived by physics knowledge and by machine learning algorithms is not always useful or necessary. In fact, we believe that there is plenty of room in the middle of these two approaches. Ignoring physics knowledge is equivalent to ignoring massive amounts of information-rich data, whereas avoiding machine learning approaches limits our toolbox to develop predictive models with quantified uncertainties. The two types of models can be complementary, in particular when it is broadly recognized that even when we know the physical model, we may not have enough resources to compute it. Hybrid approaches complementing, for example, physics knowledge with machine learning closures¹⁰, or constraining machine learning models with physics are essential. The application of uncertainty quantification techniques then becomes even more challenging, but presents an exciting scientific frontier.






F.L. Integrating physics knowledge with deep learning is key to interpretable and robust inference. The goal of the physicist is not only to build a model that can explain the data but to also do so using only a minimal set of interpretable components and parameters. Understanding the motivation for these model components and how they are causally connected is where the physics lies.

In many situations, when building a model to describe observed data, physicists have to incorporate effects that do not have

a known analytical description or that are difficult to describe from first principles. To do so, they have traditionally relied on simple empirical models that are typically ad hoc and, in some cases, may not be complex enough to accurately model the observed data, leading to systematic errors. For example, in cosmology, modelling how many galaxies are expected to exist in a dark matter halo of a given mass is traditionally modelled using empirical halo occupation distributions, which constitutes one step of the full physical forward model tying cosmological parameters to the observed galaxy distribution on the sky.

Intrinsically, these empirical components in physical models are nothing more than conditional distributions; and with the rise of efficient density estimators and generative models, it becomes possible to avoid making explicit assumptions on their analytical forms. Instead, one can use neural density estimators to model these components within a larger physical model in an agnostic and data-driven way. The parameters of these neural networks become part of the physical probabilistic model, and may be inferred from the data along with the rest of the model parameters using modern techniques such as variational inference.

The main advantage of such hybrid models in physics is the ability to retain a forward modelling approach, with a meaningful causal structure. Even if some components become empirical and data-driven, they still have a specific meaning in the larger model and remain, in that sense, interpretable.

Yarin Gal¹ , Petros Koumoutsakos , Francois Lanusse³ , Gilles Louppe⁴  and Costas Papadimitriou⁵ 


¹Oxford Applied and Theoretical Machine Learning Group, Department of Computer Science, University of Oxford, Oxford, UK.

²Computational Science and Engineering Laboratory, School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, USA.

³CNRS, CEA Saclay, Saclay, France.

⁴University of Liège, Liège, Belgium.

⁵Department of Mechanical Engineering, University of Thessaly, Volos, Greece.

 e-mail: yarin.gal@cs.ox.ac.uk; petros@seas.harvard.edu; francois.lanusse@cea.fr; g.louppe@uliege.be; costasp@uth.gr

<https://doi.org/10.1038/s42254-022-00498-4>

Published online: 22 August 2022

1. Cranmer, K., Brehmer, J. & Louppe, G. The frontier of simulation-based inference. *Proc. Natl Acad. Sci. USA* **117**, 30055–30062 (2020).
2. Abdar, M. et al. A review of uncertainty quantification in deep learning: techniques, applications and challenges. *Inf. Fusion* **76**, 243–297 (2021).
3. Hüllermeier, E. & Waegeman, W. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Mach. Learn.* **110**, 457–506 (2021).
4. Gal, Y. & Ghahramani, Z. Dropout as a Bayesian approximation: representing model uncertainty in deep learning. In *Proc. 35th International Conference on International Conference on Machine Learning* Vol 48 1050–1059 (PMLR, 2016).
5. Neal, R. M. *Bayesian Learning for Neural Networks* (Springer, 1996).
6. van Amersfoort, J., Smith, L., Teh, Y. W. & Gal, Y. Uncertainty estimation using a single deep deterministic neural network. In *Proc. 37th International Conference on Machine Learning* Vol 119 9690–9700 (PMLR, 2020).
7. Liu, J. Z. et al. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. In *Proc. 34th International Conference on Neural Information Processing Systems (NIPS'20)* 7498–7512 (Curran Associates, 2020).
8. van Amersfoort, J., Smith, L., Jesson, A., Key, O. & Gal, Y. Improving deterministic uncertainty estimation in deep learning for classification and regression. Preprint at <https://arxiv.org/abs/2102.11409v1> (2021).
9. Martin, S. M. et al. Korali: efficient and scalable software framework for Bayesian uncertainty quantification and stochastic optimization. *Comput. Methods Appl. Mech. Eng.* **389**, 114264 (2022).
10. Bae, H. J. & Koumoutsakos, P. Scientific multi-agent reinforcement learning for wall-models of turbulent flows. *Nat. Commun.* **13**, 1443 (2022).
11. Berger, J. O. & Smith, L. A. On the statistical formalism of uncertainty quantification. *Annu. Rev. Stat. Appl.* **6**, 433–460 (2019).

Acknowledgements

Y.G. holds a Turing Artificial Intelligence Fellowship at the Alan Turing Institute, which is supported by Engineering and Physical Sciences Research Council (EPSRC) grant reference V030302/1.

Author contributions

All authors contributed equally to the preparation of this manuscript.

Competing interests

The authors declare no competing interests.

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© Springer Nature Limited 2022