PHILOSOPHICAL TRANSACTIONS A

royalsocietypublishing.org/journal/rsta

Research

Article submitted to journal

Subject Areas:
Machine Learning, Uncertainty Quantification

Keywords:
Conformal Prediction, Trustworthy AI, Reliable Machine Learning

Author for correspondence:
Khuong An Nguyen
e-mail: Khuong.Nguyen@rhul.ac.uk

Preface to 'Advancing uncertainty quantification in Artificial Intelligence systems using Conformal Prediction'

Khuong An Nguyen[1], Rina Foygel Barber[2], Vicky Copley[3], Alex Gammerman[1], Vladimir Vovk[1], and Johanna Ziegel[4]

[1]Royal Holloway University of London, Surrey, UK
[2]University of Chicago, Illinois, USA
[3]Defence Science and Technology Laboratory, Ministry of Defence, UK
[4]ETH Zurich, Switzerland

As Artificial Intelligence (AI) systems are being widely deployed in safety-critical and high-stakes applications (e.g. medical diagnosis, autonomous vehicles, financial risk assessment), there is a growing demand for providing reliable and trustworthy machine predictions. However, since AI models become more complex in structure (a prominent example being Deep Neural Networks) and bigger in size (e.g. Large Language Model systems such as ChatGPT and Gemini), being able to understand, explain, and quantify confidence in their predictions are on-going challenges.

Therefore, this special issue is dedicated to exploring the forefront of reliable uncertainty quantification in AI systems, using Conformal Prediction (CP), a leading statistical framework that offers predictions with valid coverage guarantees under minimal assumptions. The issue comprises the most recent, most novel and most practical developments of CP-based methods in cutting-edge AI applications, highlighting improvements over traditional methods.

THE ROYAL SOCIETY PUBLISHING

1. A Brief Introduction to Conformal Prediction

Conformal Prediction (CP) is a general framework for quantifying uncertainty in machine learning and statistics. Its most valuable feature is that it provides distribution-free guarantees for the quality of its predictions, i.e., guarantees that do not require strong assumptions on the form of the data-generating distribution, such as a correctly specified parametric model.

In this section we will briefly review CP and related methods, including areas represented by the articles in this special issue but also taking a wider approach. The readers who are interested in further details can consult several available extensive reviews. The first book-length account was the monograph "Algorithmic Learning in a Random World" [1] by Vovk, Gammerman and Shafer, originally published in 2005 and now in its greatly expanded second edition (2022). It remains a wide-ranging account of CP and related methods. The 2008 survey by Shafer and Vovk "A Tutorial on Conformal Prediction" [2] remains popular. More recently, Angelopoulos and Bates published their "Conformal Prediction: A Gentle Introduction" [3], an implementation-oriented treatment of CP. We especially recommend the forthcoming textbook "Theoretical Foundations of Conformal Prediction" by Angelopoulos, Barber, and Bates [4], which covers a plethora of topics in a rigorous and accessible manner. Together, these publications provide a strong foundation that will allow the reader to engage with this rapidly expanding field.

The most standard form of CP converts outputs of a machine-learning or statistical model into prediction sets with a user-specified error rate. Conformal methods use the empirical behaviour of the prediction errors or more general conformity scores to calibrate uncertainty around a base prediction algorithm. This makes CP particularly appealing in modern settings where black-box models cannot deliver uncertainty estimates. We can combine conformal calibration with any machine-learning and statistical models, such as neural networks, tree ensembles, quantile regression, probabilistic predictors, as long as suitable calibration scores can be defined. This flexibility explains the growing prominence of conformal prediction across trustworthy AI.

CP has been developed for both classification and regression problems, but its modern extensions also include multi-label outputs and other kinds of complex labels. A pioneering paper developing conformal regression was published by Lei and Wasserman [5,6]; it established powerful results about asymptotic efficiency of conformal regression. A major advance was the introduction of conformal predictors based on quantile regression by Romano et al. [7]; this gives highly adaptive prediction intervals that have become very popular. Multi-label learning is reviewed in the paper by Papadopoulos in this special issue.

The basic theory of CP centres on the assumption of exchangeable data. The most radical deviation from exchangeability happens in time series, and there have been quite a few developments of CP even in this area. An early paper providing interesting theoretical results in this setting was published by Chernozhukov et al. in 2018 [8]. However, it appears that the most fertile area in the current adaptations of CP to non-exchangeable data is where exchangeability is violated but not wiped out completely. Adaptive conformal prediction introduced by Gibbs and Candès [9] has a guaranteed property of validity without any assumptions about the data but can be expected to be efficient when deviations from exchangeability are moderate, such as under a slow distribution shift. An especially interesting development of CP has been accomplished under covariate shift, the kind of distribution shift where the conditional distribution of the labels given the object features does not change. Tibshirani et al. [10] developed an ingenious but very natural weighted version of CP that assigns different weights to different training objects. Podkopaev and Ramdas [11] adapted these ideas to the symmetric situation of the label shift, where the conditional distribution of the objects given the labels does not change. An interesting further development was by Barber et al. [12], who do not make any assumptions about the kind of the distribution shift apart from the violation of exchangeability being small, in some sense.

Weighted conformal prediction gave rise to new approaches in causal inference. Under the simplest version of the popular Rubin causal model (going back to Neyman), each unit (such as a patient in a hospital) is assumed to have two potential outcomes, one corresponding to

being treated and the other to being left untreated. Both potential outcomes are hidden at first, but then one of them is revealed according to whether the unit has been treated or not. The distribution of the features may be very different between treated and untreated units, and predicting the outcome for a new unit based on a training set of treated and untreated units is akin to prediction under covariate shift and can be done effectively using weighted CP. This idea has been implemented by Lei and Candès [13] and further developed by [14]. Another approach to CP under the Rubin causal model is due to Chernozhukov et al. [15], who design ways of evaluating effects of policy interventions both under exchangeability and for time series.

Conformal risk control, introduced by Angelopoulos et al. [16], is another useful generalisation of CP. It is a step towards decision making; namely, conformal risk control presents CP as a binary decision making problem (with the loss function being the indicator function of the failure to cover the true label) and then generalises it to a wide range of loss function. This covers, e.g., false negative rate in image segmentation and multilabel classification. One of the articles in this special issue by Angelopoulos generalises conformal risk control to an even wider class of loss functions.

The main property of validity of CP is that, when its predictions are presented as prediction sets, their coverage probability is equal to the desired target error rate. This probability, however, is marginal. Marginal coverage is a necessary criterion in most applications. However, in many situations, it is desirable to have conditional coverage guarantees of various forms. Natural possible choices are conditioning on the training set, on the test object, and on the test label; the most popular kind of conditional validity has been object-conditional. There have been many interesting negative results showing the limits of conditional inference for any methods of prediction with certain guarantees of validity, not just CP. Important milestones here are [6] and [17]. The recent paper by Gibbs et al. [18] defines a spectrum of problems that interpolate between marginal and conditional validity.

The most basic version of CP, often referred to as full CP, is computationally efficient only for a narrow range of base prediction algorithms, those that have particularly convenient mathematical descriptions, such as ridge regression or lasso [19]. In most interesting cases, such as that of deep neural networks, full CP is computationally infeasible. The most straightforward way of achieving computational efficiency is to split the training set into two parts, use one part only for training the base algorithm, and use the other part for turning point predictions into prediction sets. This is known as inductive CP or split CP. The disadvantage of this method is that it loses some predictive efficiency as compared with full CP, since the latter uses the full training set both for training the base algorithm and for turning its predictions into prediction sets. The method of jackknife+ developed by Barber et al. [20,21] greatly improves the predictive efficiency of inductive CP; on the negative side, it somewhat weakens the guaranteed property of validity that full CP possesses.

While basic conformal prediction is a way of producing predictions sets, it can be adapted to probabilistic prediction, both in the case of regression and in the case of classification. This is reviewed in [1, Part II] and [4, Chap. 12] and further developed in a paper in this special issue [Residual Distribution Predictive Systems by Allen et al.].

The final breakthrough that we discuss in this brief survey is "conformal training" developed by a team at DeepMind [22]; independently, similar ideas were proposed in [23] and [24]. In conformal training, we train a conformal predictor directly, instead of using CP as a wrapper around a base algorithm.

The most prolific application domains of Conformal Predictors in the published literature include clinical and biomedical AI, pharmaceutical drug discovery, electoral forecasting, autonomous systems, and precision agriculture. In the pharmaceutical sector, for instance, AstraZeneca has successfully deployed Conformal and Venn Predictors (CVP) to reduce the number of compounds requiring laboratory testing, thereby accelerating discovery and lowering costs [25]. In medicine, ongoing work in breast cancer risk assessment aims to improve diagnostic quality [26]. A widely cited example outside the life sciences is the Washington Post's use

of conformal prediction to forecast ranges of likely outcomes in U.S. elections across states and county types [27]. Beyond these areas, conformal prediction underpins Microsoft's Azure anomaly-detection framework and has shown promise in precision agriculture, particularly in improving the efficiency of automated weeding systems [28]. Numerous additional applications can be explored in public technical commentary, including curated collections such as [29].

All these developments, the developments described elsewhere in this special issue, and many more not included here, demonstrate the flexibility of the original concept of CP. We hope that CP and related methods will find many more useful applications supported by new interesting theoretical results.

2. Theme Issue Structure

This theme issue brings together ten articles that complementarily address the above challenges using conformal prediction. The articles are divided in three broad thematic groups.

The first group surveys rapidly developing research areas in which uncertainty-aware methods are becoming increasingly important.

- "Uncertainty-Aware Large Language Models: A Scoping Review of Conformal Prediction Methods" by Ashby et al. surveys applications of Conformal Prediction to large language models, proposing a six-part taxonomy and analysing trends across 106 papers.
- "Concerning Uncertainty—A Systematic Survey of Uncertainty-Aware XAI" by Löfström et al. reviews how uncertainty is incorporated into explainable AI, highlighting evaluation gaps and the need for reliability-aware explanatory pipelines.
- "Conformal Prediction for Multi-Label Learning: A Review of Methods and Guarantees" by Papadopoulos consolidates conformal approaches to multi-label prediction, comparing guarantees, dependence modelling, scalability, and output structure.

The second group extends the theory and scope of conformal methods, including Bayesian reasoning and risk control.

- "Residual Distribution Predictive Systems" by Allen et al. develops an alternative route to predictive systems with out-of-sample calibration guarantees.
- "The Interplay between Bayesian Inference and Conformal Prediction" by Deliu and Liseo explores how Bayesian procedures and conformal calibration can complement one another, balancing informativeness with frequentist validity.
- "Conformal Risk Control for Non-Monotonic Losses" by Angelopoulos generalises conformal risk control beyond monotonic one-dimensional losses to non-monotonic objectives and multidimensional parameter settings.
- "CP4SBI: Local Conformal Calibration of Credible Sets in Simulation-Based Inference" by Cabezas et al. brings conformal calibration into simulation-based Bayesian inference.

The third group demonstrates how conformal prediction can support reliability in applied settings where uncertainty is highly critical.

- "Synthetic Data Assurance with Conformal Prediction" by Copley and Hiett uses conformal tools to assess the fidelity and diversity of synthetic image data.
- "Uncertainty Quantification in Train Delay Prediction with Conformal Prediction" by Luo et al. applies conformal calibration to railway delay forecasting.
- "Uncertainty Quantification Using Conformal Prediction for Mesh-Based Simulations" by Mabtoul et al. develops conformal prediction sets for graph-neural-network surrogates in computational fluid dynamics.

Acknowledgements. The Guest Editor team would like to express their gratitude to all the reviewers for their time and expertise in improving the quality of the theme issue.

References

1. Vovk V, Gammerman A, Shafer G. 2022 *Algorithmic Learning in a Random World*. Cham: Springer second edition.
2. Shafer G, Vovk V. 2008 A tutorial on conformal prediction. *Journal of Machine Learning Research* **9**, 371–421.
3. Angelopoulos AN, Bates S. 2023 Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning* **16**, 494–591.
4. Angelopoulos AN, Barber RF, Bates S. 2026 Theoretical Foundations of Conformal Prediction. Technical Report arXiv:2411.11824 [math.ST] arXiv.org e-Print archive. Pre-publication version of a book to be published by Cambridge University Press.
5. Lei J, G'Sell M, Rinaldo A, Tibshirani RJ, Wasserman L. 2018 Distribution-Free Predictive Inference for Regression. *Journal of the American Statistical Association* **113**, 1094–1111. (10.1080/01621459.2017.1307116)
6. Lei J, Wasserman L. 2014 Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society B* **76**, 71–96.
7. Romano Y, Patterson E, Candès EJ. 2019 Conformalized quantile regression. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)* pp. 3543–3553.
8. Chernozhukov V, Wüthrich K, Zhu Y. 2018 Exact and robust conformal inference methods for predictive machine learning with dependent data. *Proceedings of Machine Learning Research* **75**, 732–749. COLT 2018.
9. Gibbs I, Candès EJ. 2021 Adaptive conformal inference under distribution shift. *Advances in Neural Information Processing Systems* **34**, 1660–1672.
10. Tibshirani RJ, Barber RF, Candès EJ, Ramdas A. 2019 Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems 32* pp. 2530–2540. Curran Associates.
11. Podkopaev A, Ramdas A. 2021 Distribution-free uncertainty quantification for classification under label shift. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence* pp. 844–853.
12. Barber RF, Candès EJ, Ramdas A, Tibshirani RJ. 2023 Conformal prediction beyond exchangeability. *Annals of Statistics* **51**, 816–845.
13. Lei L, Candés EJ. 2021 Conformal inference of counterfactuals and individual treatment effects. *Journal of the Royal Statistical Society, Series B* **83**, 911–938.
14. Jin Y, Ren Z, Candès EJ. 2023 Sensitivity analysis of individual treatment effects: A robust conformal inference approach. *Proceedings of the National Academy of Sciences* **120**, e2214889120.
15. Chernozhukov V, Wüthrich K, Zhu Y. 2021 An exact and robust conformal inference method for counterfactual and synthetic controls. *Journal of the American Statistical Association* **116**, 1849–1864.
16. Angelopoulos AN, Bates S, Fisch A, Lei L, Schuster T. 2024 Conformal risk control. In *Proceedings of ICLR 2024*.
17. Barber RF, Candès EJ, Ramdas A, Tibshirani RJ. 2021 The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA* **10**, 455–482.
18. Gibbs I, Cherian JJ, Candès EJ. 2025 Conformal prediction with conditional guarantees. *Journal of the Royal Statistical Society B* **87**, 1100–1126.
19. Lei J. 2019 Fast exact conformalization of lasso using piecewise linear homotopy. *Biometrika* **106**, 749–764.
20. Vovk V. 2015 Cross-conformal predictors. *Annals of Mathematics and Artificial Intelligence* **74**, 9–28.
21. Barber RF, Candès EJ, Ramdas A, Tibshirani RJ. 2021 Predictive inference with the jackknife+. *Annals of Statistics* **49**, 486–507.
22. Stutz D, Dvijotham KD, Cemgil AT, Doucet A. 2022 Learning optimal conformal classifiers. In *Proceedings of ICLR 2022*.
23. Bellotti A. 2021 Optimized conformal classification using gradient descent approximation. Technical Report arXiv:2105.11255 [cs.LG] arXiv.org e-Print archive.
24. Colombo N, Vovk V. 2020 Training conformal predictors. *Proceedings of Machine Learning Research* **128**, 55–64. COPA 2020.

25. Toccaceli P, Nouretdinov I, Gammerman A. 2017 Conformal prediction of biological activity of chemical compounds. *Annals of Mathematics and Artificial Intelligence* **81**, 105–123. (10.1007/s10472-017-9556-8)
26. Fröhlich A, Ramos T, Santos GMCD, Buzatto IPC, Izbicki R, Tiezzi DG. 2025 PersonalizedUS: Interpretable Breast Cancer Risk Assessment with Local Coverage Uncertainty Quantification. In *Proceedings of the AAAI Conference on Artificial Intelligence* vol. 39 pp. 27998–28006. (10.1609/aaai.v39i27.35017)
27. Cherian J, Bronner L. 2020 How The Washington Post Estimates Outstanding Votes for the 2020 Presidential Election. Technical report.
28. Melki P, Bombrun L, Diallo B, Dias J, da Costa JP. 2025 Uncertainty Guarantees on Automated Precision Weeding using Conformal Prediction. (10.48550/arXiv.2501.07185)
29. Towards Data Science. 2026 All You Need Is Conformal Prediction. https://towardsdatascience.com/?s=All+You+Need+Is+Conformal+Prediction. Accessed: 2026-05-29.