Personalized uncertainty quantification in artificial intelligence

Article

Accepted Version

Personalised Uncertainty Quantification

Tapabrata Chakraborti[1,2,*], Christopher R.S. Banerji[1,2,3], Ariane Marandon[1], Vicky Hellon[1], Robin Mitra[2], Brieuc Lehmann[2], Leandra Bräuninger[2], Sarah McGough[4], Cagatay Turkay[5], Alejandro Frangi[6], Ginestra Bianconi[7], Weizi Li[8], Owen Rackham[9], Deepak Parashar[1,5], Chris Harbron[10], and Ben D. MacArthur[1,9]

[1]The Alan Turing Institute, London, UK
[2]University College London, London, UK
[3]King's College London, London, UK
[4]Genentech, South San Francisco, CA, USA
[5]University of Warwick, Coventry, UK
[6]University of Manchester, Manchester, UK
[7]Queen Mary University, London, UK
[8]Henley Business School, University of Reading, UK
[9]University of Southampton, Southampton, UK
[10]Roche Pharmaceuticals, Welwyn Garden City, UK
[*]Correspondence to: Tapabrata Chakraborti (tchakraborty@turing.ac.uk; t.chakraborty@ucl.ac.uk)

ABSTRACT

Artificial Intelligence (AI) tools are increasingly being used to help make consequential decisions about individuals. While AI models may be accurate on average, they can simultaneously be highly uncertain about outcomes associated with specific individuals or groups of individuals. For high stakes applications (such as healthcare and medicine, defence and security, banking and finance), AI decision support systems must be able to make personalised assessments of uncertainty in a rigorous manner. However, the statistical frameworks needed to do so are currently incomplete. Here, we outline current approaches to Personalised Uncertainty Quantification (PUQ) and define a set of grand challenges associated with the development and use of PUQ in a range of areas, including multimodal AI, explainable AI, generative AI, and AI fairness.

1 Introduction

Artificial Intelligence (AI) based predictive models are starting to find use in highly sensitive systems – such as healthcare and precision medicine [3], defense and criminal profiling [15], banking and financial forecasting [24] – in which decisions of high consequence are made about individuals. While AI models may be accurate at a population level (i.e., on average), they can simultaneously be uncertain about certain individuals or groupings of individuals. For high stakes applications, such decisions systems must be able to provide personalised assessments of uncertainty in a rigorous manner. However, model uncertainty will typically depend on many factors, such as protected characteristics and other demographic information (that, in turn, are associated with data availability/missingness) and is therefore highly personalised. Failing to understand such personalised uncertainty can hide significant shortcomings in predictive model performance.

Typically, uncertainty is viewed as deriving either from (1) the data used to train the model (in which case, it is referred to as being aleatoric) or (2) the model itself (in which case, it is referred to as being epistemic). Aleatoric uncertainty (or statistical uncertainty) arises due to inherent randomness or measurement noise, and cannot be reduced. Epistemic uncertainty

(or systematic uncertainty) arises due to lack of knowledge or limited data, and can be reduced with more information or better models [45]. For example, weather forecasting exhibits aleatoric uncertainty due to chaotic atmospheric dynamics; poor weather models also have epistemic uncertainty, that can be improved with better training data and methods. Similarly, in a medical diagnosis context, aleatoric uncertainty may arise from the inherent variability in patient responses to a treatment, while epistemic uncertainty may arise due to incomplete knowledge about the relationships between symptoms and diseases. Often, it is hard to disentangle sources of uncertainty and these two classes are not always distinguished in practice [28]. Moreover, epistemic and aleatoric uncertainties may be interrelated. For instance, including additional predictors (i.e., increasing the input data dimensionality) may reduce aleatoric uncertainty at the expense of increased epistemic uncertainty [45]. This interrelation can make it challenging to marry traditional statistical concepts of uncertainty with the epistemic/aleatoric dichotomy, because notions such as the 'p-value' typically quantify uncertainty deriving from both data and model. More generally, uncertainty quantification concepts are often encountered indirectly in accuracy vs. precision tradeoffs [85] (also referred to as bias-variance trade-offs in the statistical literature [88]). Accuracy quantifies how close a prediction is to the true value, while precision quantifies the reproducibility of repeated predictions, with the same expected outcome. Sources of uncertainty can thus be categorised by their impact on either the accuracy or the precision of the model prediction. As with the epistemic/aleatoric dichotomy, precision and accuracy are interrelated, for example decreasing the precision of a highly accurate model, will also generally lower its accuracy [45]. Moreover, again, they refer population-level, rather than individualized, metrics.

Tools that are tailored to individual predictions and are an inherent part of the machine learning process are now needed. However, assessments of personalised uncertainty are rarely used as performance metrics when training machine learning models. Rather, the performance of a predictive model is usually assessed by its accuracy, typically using a single metric that summarises the model performance over the whole of a held-out test dataset – for example the proportion of correct predictions or an Area Under the Curve (AUC) assessment. Sometimes such metrcs are reported per class, and additional accuracy related metrics like precision, recall or F1 score may also be used [53]. However, the uncertainty associated with individual predictions may vary considerably, and so even in a mostly accurate model, some individual predictions may be poor. This can happen due to a combination of many factors such as a lack of similar training data points for some individuals, a greater proportion of missing data in some regions of the training data, inappropriate extrapolations being made either explicitly or implicitly within the model, or some individuals being edge cases and so fundamentally harder to predict. For a model to support robust decision-making in high-stakes contexts it should account for predictive uncertainty on an individualised basis and, moreover, be able to incorporate this information into the decision-making process.

Doing this is a significant challenge that we have yet to solve properly. For this reason, the Alan Turing Institute (UK's national institute for AI), in partnership with Roche Pharma, hosted a series of workshops in 2022-2023 to convene a multidisciplinary community of experts to identify the outstanding problems in personalized uncertainty quantification (PUQ), and sketch a road map for their solution. This paper is the fruit of those meetings. The article is organized as follows. Section 2 provide an introduction to PUQ within the wider statistics and machine learning literature, with a particular focus on conformal prediction as a particularly promising methodology, among others. In Section 3 we then outline a set of grand challenges in PUQ, which if resolved will substantially advance the field, and place us in a stronger position to appropriately deploy machine learning methods. Section 4 will end with some forward-looking concluding remarks.

2 Towards Personalised Uncertainty Quantification

Consider the situation in which an AI-based clinical decision support system helps provide a patient prognosis for a clinician. Suppose for a particular patient, it predicts that they do not have a given disease, but also has a high level of uncertainty in that prediction. This is equivalent to saying that the patient has unknown prognosis (or all possible prognoses, including poor outcomes) and so is of no practical use to the clinician. Moreover, it is less reassuring to the patient than a prediction

of a moderate outcome with high certainty since this would rule out particularly poor outcomes (and would also, in turn, avoid the need for invasive or burdensome further testing). In practice, such situations may be common since the predictive uncertainty of the model may vary considerably between patients, reflective of many factors including the level of atypicality of the individual, the number of similar cases in the training data set used to derive the model, or the lack of availability of some data modalities, such as genetic testing, for some patients [70]. These differences may also reflect the rarity of some conditions, and associated lack of knowledge, or the fact that less data was collected for some patient populations (which could be associated with demographic factors, such as deprivation) than others [66].

For a prediction to have utility for decision making, then, the prediction point estimate must be placed into context, including an understanding of the uncertainty attached to it. Often this takes the form of a prediction interval that provides a range of values in which the yet to be observed value will fall with a specified probability [81]. Within traditional statistical models based on, for example, linear or logistic regression, there is a well-established theory to generate such prediction intervals. However, these methods do not automatically extend to more general data driven modelling scenarios, particularly when data distribution assumptions are violated. Non-parametric techniques such as bootstrapping, which can theoretically be applied to any modelling technique, may be useful in these situations but are typically computationally expensive to implement [54]. Similarly, Bayesian modelling frameworks [35] naturally allow for the generation of credible intervals from the posterior distribution, whilst notions such as Monte Carlo (MC) dropout can be used to estimate uncertainty in deep leading models via generating an appropriate ensemble [61]. However, each of the different methods described above rests on different assumptions, and thus results in different uncertainty intervals, even if they ostensibly have the same coverage, implying that some intervals may represent weaker or less representative bounds on the 'true' model uncertainty.

A number of recent methodological advancements have started to address this challenge,. For example, a Dirichlet-based method was proposed to quantify personalised uncertainty across heterogeneous data sources in federated learning scenarios [52]. By integrating improved posterior networks, the methodology adjusts predictions for individual data distributions while estimating their uncertainty. Similarly Bayesian deep learning based approaches for quantifying uncertainty of personalised recommendations have also been suggested – for example, in the context of collaborative filters which are popular in recommender systems [89]. Though these two methods make useful advancements in quantification of personalised uncertainty, they do so in very specific settings (federated learning and recommender systems respectively). Yet, there is a need to go beyond such specific situations and formulate principled ways to establish rigorous bounds independently of the training data distribution and model architecture, and can be generalised across most application areas.

To meet this challenge, conformal prediction (CP) [86] is a modern uncertainty quantification technique, that is applicable to any AI algorithm, which has garnered significant interest in the recent years [7]. In contrast to other uncertainty quantification methods, CP provides guarantees in real-world scenarios (albeit with some caveats, as described below) subject to minimal assumptions regarding sample size and data distribution. We next present CP and then proceed to describe how it represents a powerful tool for moving towards PUQ.

2.1 Conformal prediction for PUQ

At a high-level, conformal prediction (CP) [86, 65, 7], is a technique that converts the predictions of an arbitrary AI algorithm into set-valued predictions, called *prediction sets*, that contain the true outcome with a probability above a user-specified level $1-\alpha$. The general idea underlying CP is to use a hold-out dataset, known as the calibration set, that has not been used for training the predictive model, to rigorously assess the uncertainty of the model's predictions.

To see how this works, let X denote a covariate variable, such as a vector of unidimensional measures, or an image, and Y denote an outcome variable, which may be categorical or continuous. Let $(X_i, Y_i)_{i=1}^{n}$ be an exchangeable hold-out sample of observations and X_{n+1} be a new patient for which we would like to predict Y_{n+1} (the new observation also being exchangeable).

An assumption of conformal prediction is that the sequence of data points (hold-out examples and new test points) is

exchangeable. This means that the probability distribution of the data does not change when the order of the data points is altered. Exchangeability is important for the validity of conformal prediction because the method relies on ranking or comparing the conformity scores of data points. Formally, the general aim of conformal prediction is to build the smallest possible prediction set $\hat{C}_\alpha(X_{n+1})$ such that the probability of the true outcome being in the prediction set (*coverage*) is above $1-\alpha$. That is,

$$\mathbb{P}(Y_{n+1} \in \hat{C}_\alpha(X_{n+1})) \geq 1-\alpha,$$

where in the above equation, the probability is over $(X_i, Y_i)_{i=1}^{n+1}$, and α is a user-specified margin of error (e.g., 5%). The above guarantee has a concrete, intuitive interpretation: for $1-\alpha$% of data points, the output prediction sets will contain the true outcome.

To achieve a coverage above the target confidence level, CP relies on the use of a *non-conformity score* function $\nu : (x, y) \mapsto \mathbb{R}$, that measures the inadequacy of y as a prediction for x based on a pre-trained predictive model. In a regression task, a classical choice for the non-conformity score is $\nu(x, y) = |\hat{\mu}_x - y|$, where $\hat{\mu}_x$ is the point prediction given by a trained regressor for the data point x. In the classification case, the non-conformity score is typically $\nu(x, y) = 1 - \hat{p}_y(x)$, where $\hat{p}_y(x)$ is the estimated probability that the data point x is in class y given by a trained classifier.

For a given non-conformity score, CP constructs a prediction set for Y_{n+1} by comparing the non-conformity score of (X_{n+1}, y) for any possible outcome y to the non-conformity scores on the hold-out *calibration* data: y is excluded from the prediction set $\hat{C}_\alpha(X_{n+1})$ only if the non-conformity score for y is large enough with regards to this set of values, that is, larger than the $(1-\alpha)\frac{n+1}{n}$ quantile. Hence, the hold-out data is used to build a reference set representing the predictive model's estimated uncertainty associated to true outcomes. Moreover, under exchangeability of $(X_i, Y_i)_{i=1}^{n+1}$, CP guarantees that the coverage is above $1-\alpha$. Crucially, this guarantee holds regardless of the size n of the hold-out set, the type of data distribution (as long as exchangeable), the predictive model used and its accuracy. At the same time, the larger the hold-out sample size and the more accurate the predictive model, the more informative or smaller the prediction sets will be.

While strict exchangeability is a standard assumption, there are methods to adapt conformal prediction to settings where exchangeability is not fully met, such as when dealing with time-series data or other structures with complex dependence [13]. One such method is online conformal inference [38, 39, 5, 91]. In online settings data are usually collected in sequence and the task is to construct a prediction set at each time step for a sequence of pairs of data points (X_i, Y_i), with the size of the set determined by a data-driven threshold for the non-conformity score. After observing the corresponding outcome label Y_i, these methods update the threshold for the non-conformity score depending on whether the algorithm has been under/over-covering at previous timesteps for the desired error rate α.

While conformal prediction sets are specific to each individual, it is important to emphasize that the aforementioned guarantee is only *marginal*: it ensures that the average coverage is above the confidence level, but does not ensure uniform coverage across the feature space and the outcome space. In other words, it may occur that on certain parts of either the feature space or the outcome space, the coverage is below the level of confidence (i.e, under-covered), while it is above it on the remaining parts of the space (i.e., over-covered). For instance, parts of the feature/outcome space which are under-represented in the hold-out sample are more likely to be under-covered [55, 57]. Alternatively, if there is heteroscedasticity in the outcome's distribution given the covariates then parts of the feature space where the variance of the outcome is larger are also more likely to be under-covered. Thus, while CP provides a powerful framework for assessing PUQ it is not yet complete. Indeed, a range of challenges still remain.

3 Grand challenges in PUQ

To scope these challenges, the Alan Turing Institute convened a series of workshops to understand the issues and determine a trajectory for future research in CP and PUQ more generally. In this section we outline eight grand challenges in PUQ, that

arose from these discussions. Collectively, they scope the problems associated with PUQ, provide a road map for the further development of PUQ as a field of study, and outline a set of future research directions that will help advance PUQ for machine learning at scale. These grand challenges are also visually summarised in Figure 2.

3.1 PUQ for individualised predictions

In the last section, we discussed several methods for PUQ, including CP. We noted that although CP is powerful, it only provides marginal guarantees; yet, truly personalized uncertainty quantification requires guarantees conditioned on the individual, since they ensure that the probability of error is the same for all individuals. While exact conditional coverage is known to be impossible without distributional assumptions [33], various approaches have been proposed to tackle this issue in practice. For example, heuristics for improving conditional coverage empirically, either through a special choice of the non-conformity score [72, 73] or a modification of the conformal procedure itself [42, 29] have been proposed. A special case of interest consists of aiming for equal coverage across a finite number of non-overlapping groups, such as age groups. In this case, a standard solution is to use Mondrian CP (MCP), a minor extension of classical CP in which calibration is performed separately for each group [86, 71, 75]. MCP achieves equal group-wise coverage subject to the trade-off that prediction sets are constructed using a smaller number of calibration samples. Recently, a relaxation of the conditional coverage objective and generalization of the CP procedure which achieves exact guarantees (and therefore includes group-conditional coverage as a special case) has been proposed [40]. Such innovations are a step towards achieving PUQ, yet a general solution is elusive, particularly when we seek to strengthen PUQ from sub-group level guarantees to the level of specific individuals, which is the ultimate goal in this area.

3.2 PUQ for multi-scale modelling

Multi-scale modelling involves the integration of information at different length or timescales (e.g., from microscopic to macroscopic levels) or resolutions (e.g., combining individual-level with population-level information). Incorporating data at different scales not only increases model complexity, but also poses challenges to quantifying personalised uncertainty [90]. These challenges often arise from one of three sources: scale integration, data diversity/quality, and model complexity/validation [2]. Issues that arise from scale integration are related to lack of understanding of the non-linearities associated with moving between scales. For example, constructing a non-conformity function for conformal predictions that accounts for all scales in parallel is often intractable and, as a result, is often tackled by combining the output of scale-specific functions that do not properly account for relationships between scales [6]. Issues to do with data diversity or quality arise because data heterogeneity will often differ considerably between scales. For instance, a disease risk prediction model that incorporates environmental information at the postcode level with genetic testing at the individual level will have to account for very different sources of hetereogeneity. Genetic data will be highly sparse – and missing for many individuals – but, where available, highly accurate; whilst environmental data may be mostly complete but much coarser. In both cases, underlying biases pose different challenges both in ensuring accuracy and in quantifying uncertainties associated with individual predictions [1] – these challenges are closely related to those associated with PUQ for multimodal data (see Challenge 7). Accurately assessing personalised uncertainties in such situations requires accounting for scale-dependent data quality issues. This area can be overlapping in scope with the next challenge in multimodal AI, but there is a subtle distinction which might be best illustrated with the following illustrative example from computational healthcare. When we are dealing with radiology imaging for say lung cancer we could look at the whole image in the context of the location of the nodule within the lung cavity, or we could zoom in onto the nodule and look at it morphological characteristics, each would provide different sets of information and challenges for the AI system. Same holds true when looking at histopathology image of the same nodule, where the whole slide image would give tissue level information and resolutions whereas the most zoomed in view would provide cellular level tumour microenvironment information and resolutions. In each of these cases the modality remains the same, either radiology or pathology. However if we combine the two into the same decision pipeline, plus perhaps include further modalities like

clinicogenomcis, then the system becomes multimodal and brings with it a whole new set of challenges regarding compatibility of data structures, information overlap or orthogonality or redundancy, etc. Thus a grand challenge in this area of multi-scale modelling is to design non-conformity functions such that they are robust to data qualities across different scales.

3.3 PUQ for multimodal AI

As technology advances, new and varied data modalities are emerging that provide complementary sources of information. For instance, the growth of -omics approaches to obtain biomedical data has brought the desire to integrate diverse biological data sources (e.g., genomics, proteomics, metabolomics, transcriptomics, radiomics etc) [50, 48]. The increasing availability of wearable sensors and digital health solutions promises to bridge these insights to a genuinely dynamic understanding of health and disease [21], whether at a single timepoint or through the patient's life course. These exciting developments provide rich inputs to inform predictive models but also beg the question of whether less is more. Intuitively, the more inputs into a predictive model, the better informed it will be; yet, introduction of unnecessary confounders increases the risk that nuisance variables, which increase uncertainty, will be included. The challenge is thus to develop frameworks that account for the trade-off between increased information and uncertainties. A central insight to this trade-off comes from the value of information theory: a predictive model is as good as its ability to enable decision-making, bounded by the data and time available and accounting for associated uncertainties [34]. A key challenge is thus to assess models and uncertainty in a way that considers additional data and data modalities and their impact on decision-making at a individual level. In some cases additional data or modalities may not necessarily be beneficial and the benefits derived depend on the specific context of the use of the predictive model. In addition to changes in the inter-relations in a given data set, new data modalities to augment the existing model could also become available over the life-cycle of a predictive model. While new data sources can offer promise to improve model performance and address data scarcity, they bring with them new sources of uncertainties and questions about how to best integrate then into existing model designs [19]. The emerging field of *multimodal learning* offers promising solutions for such situations [10]. In these approaches, data sources that represent different aspects of the same phenomenon (i.e., modalities) can be modelled separately and/or 'fused' into a joint model. Moreover, *multimodal co-learning methods* [69] can be used to transfer learnt concepts between models, and thereby help mitigate the uncertainties associated with the introduction of new modalities into existing systems. The grand challenge in this area is to design PUQ tools that are attuned to multimodal AI methods, able to adapt to new modalities as they are incorporated, and able to quantify the predictive uncertainty for individuals for whom not all modalities are available.

3.4 PUQ for explainable AI design

Explainable AI (XAI) is the design of AI algorithms that can be interpreted or interrogated to understand why they give the output that they do for any individual [58]. Thus, XAI aims to increase trust in the predictions of a model by understanding the rationale underlying a prediction for an individual, while PUQ assessments aim to increase trust by better understanding how much confidence should be placed in that prediction [9]. This is especially true for domain inspired explainable design of AI systems [83], where the alignment to an established decision logic increases certainty and hence trustworthiness. This then links to personalised uncertainty quantification as Explainable AI may also give an extra indication of uncertainty which may not otherwise be identified if the explanation for an individual's prediction would not make sense for that individual [81]. An example of this is the combination of self-explaining models (such as concept bottlenecks or prototype based models) with conformal analysis [67]. A further interesting potential area of research could be at the intersection of explainable AI and conformal prediction, where as well as understanding the reasons for the AI model predicting the most likely output, explainable AI is also applied to understand the reasons underlying all members of the conformal prediction set. For example applied to digital pathology, this may identify a most likely pathology being driven by features observed within a certain set of cells, but also recognise there are alternative potential pathologies which would be driven by features observed in other cell

types [4]. Thus the grand challenge in this area is to employ PUQ in such a manner that it adds interpretability to personalised decisions, which by default might not be transparent to the non-specialist user.

3.5 PUQ for monitoring models

Dataset drift/shift (the gradual/sudden change in the association between model input and output over time), has the capacity to drastically limit the performance of a model trained on 'out-of-date' data [68]. In some situations, uncertainty quantification can be used to monitor the performance of predictive models over time, and identify the earliest opportunity to address performance loss due to dataset drift/shift. There are a number of ways to do this. For example, via data distribution based uncertainty monitoring [63, 36] (although this can can potentially lead to false alarms) or performance-based monitoring, which traces deviations in the model output error [77] to detect model changes and uncertainties [76]. Although useful, both these methods are population based and so may not provide the coverage guarantees at the individual level needed for high stakes applications like healthcare or criminal justice, in which individuals cannot be treated as outliers or 'out of distribution'. The main challenge is that performance-based monitoring requires rapid feedback on predictions, which is not always available in the real-world environment [36, 16], although various (model-agnostic, Bayesian and non-Bayesian) UQ tools can be used to monitor and detect model degradation and alleviate the limitations of data distribution-based and performance-based approaches [37, 54, 12, 82]. Yet, as shown in comparison studies [26, 14], no single uncertainty quantification methods works best in all data drift scenarios. Ensemble UQ tools could provide a solution by aggregating multiple uncertainty quantification methods [16]. The quality of uncertainty quantification under data shift is another research challenge as uncertainty estimation consistently degrades with increasing dataset shift regardless of method [64]. Incremental and online learning paradigms could be leveraged in model uncertainty monitoring methods, which have high capabilities in continuously adapting to accommodate the incoming data points in drift handling systems [56, 22]. Recently, some progress has been made in this area via online conformal prediction [38, 39, 5], as discussed briefly in Section 2.1, but work in this area is still nascent. Thus, there are several issues to be addressed, but the grand challenge is that the current methodologies like data distribution and performance based uncertainty monitoring do not provide PUQ; calibration based methods can but they usually need offline calibration whereas monitoring systems are mostly real-time.

3.6 PUQ with missing data

As large datasets expand over time, not only can the distribution of the data change, but also increasing amounts of missing data are likely to be present, including structured missing (SM) data in which the missingness is not random but rather exhibits some multivariate patterns of association [62]. Uncertainty quantification is strongly affected by the presence of SM, so it needs to be modelled robustly with predictive algorithms handling this problem by design. A new taxonomy for SM has recently been set out [47] providing important insights into the phenomenon that could facilitate strategies to accommodate SM into statistical modelling. However, approaching this using standard statistical techniques alone is challenging due to the ever-increasing volume and heterogeneity of large databases. In particular, the inherent pattern or structure in the missing data might be connected to certain characteristics of individuals and so must be treated accordingly in the context of PUQ. In this context, network-based approaches to understanding the geometry of missing readings provide a key complementary tool, particularly in addressing scalability issues. Networks and their generalisations, including multilayer [17] and higher-order networks [18] permit a deeper understanding from structured data. When combined with network embedding strategies [11], and information theory tools [43] they can capture the geometry and topology of missingness patterns, revealing underlying structures, which can inform the development of statistical models to be used for prediction and uncertainty quantification when making individual-level predictions. The grand challenge in this area is that the data associated with certain sub-groups of individuals may be systematically missing in specific data fields that are important to predictions, and so models built on this data will naturally be more uncertain about these groups. New tools that are able to assess the extent of this problem a priori

and flag it are needed, as a start. In the longer term tools to deal with such informative missingness will be required.

3.7 PUQ for equitable decisions

Uncertainty in personalised predictive modelling can stem from various sources. For under-represented populations, the paucity of relevant training data naturally plays a critical role: all else being equal, parameter estimates and model fits will be more uncertain for subgroups with fewer individuals. Because levels of data availability may vary with certain protected demographic characteristics like age, ethnicity, gender, [51, 84], this may increase uncertainty for those protected groups for which less information is available to predict the outcome of interest [62]. Moreover, the use of such attributes in predictive modelling may have wider societal impacts with respect to inequitable outcomes across population sub-groups, like unfair criminal profiling of racial minorities [8], which must also be taken into account [87]. Thus, increased uncertainty for under-represented populations can have significant negative consequences, and the potential adverse effects of data poverty have accordingly been well documented [46]. Polygenic risk scores (PRS) - disease prediction models based on genomic data - provide a striking example of this. The vast majority (86%, currently) of data in genomic studies comes from European-ancestry individuals [32]. Correspondingly, PRS exhibit significantly poorer performance for non-European-ancestry individuals, which, should they be used, may naturally lead to poorer clinical decisions [59]. Similar disparities have been exposed across a range of other domains, from facial recognition [20] to chest X-ray pathology discrimination [78]. Several strategies are available to mitigate such consequences. First and perhaps foremost, targeted data collection to redress imbalances in training data must be a priority [23]. In genomics, for example, significant efforts are underway to sequence the genomes of African-ancestry individuals [60]. The grouping together, or not, of individuals at the training stage can affect downstream predictive performance [80]; appropriate borrowing of information across subgroups, for example by hierarchical modelling, has the potential to boost predictive power for under-represented populations. Finally, reports of uncertainty in human-in-the-loop systems may be effective in flagging highly uncertain predictions to the eventual decision-maker, which may avert poor decisions based on insufficient information [74]. As AI models are expected to be employed increasingly in high impact decisions for individuals, the grand challenge here is to implement PUQ so that it operates fairly across population sub-groups and individuals, particularly with respect to protected attributes like race and gender.

3.8 PUQ for generative AI

The recent emergence of generative AI creates a new range of challenges for quantifying uncertainty [25], such as the worrying propensity for large language models (LLMs) [31] to make highly plausible but factually incorrect statements with an apparently high degree of confidence. Unlike outputs from more classical AI, where the prediction may be a numeric value or one of a set of categories and so an uncertainty can naturally be expressed by a range of values or by a set of categories representing a likely range of values, the best way to express uncertainty for an image, video or text document generated by AI is less clear [49]. With most classically predictive AI models, greater accuracy or decreased uncertainty as represented by a smaller prediction interval or set is desirable. However for generative AI uncertainty is context specific – for applications where accuracy is valuable, such as summarising a long document or writing an essay on a historical character, low uncertainty is valued. However when Generative AI is used in a creative scenario, for example to generate artwork, then a higher degree of variability may actually be desirable in order to create a rich and diverse portfolio of pictures [92]. Even in the scenarios valuing accuracy, lower uncertainty may be more important for some parts of the output than others. For example “Neil Armstrong was born in 1930” and “1930 is the birth year of the first man to walk on the moon” are different sentences, but the important information that both sentences convey is identical. This suggests that one approach would be to identify key features associated with a generative AI output and assess the uncertainty associated with those features, where methods similar to conformal prediction could be applied. In addition to the scale and complexity of generative AI models, the above issues make quantifying uncertainty from these models challenging. As the capability and use of generative AI models in a wide variety of domains continues to grow,

this is an area requiring greater research. Incorporating recognition of uncertainty into LLMs – changing the language from one which is frequently of certainty whether founded or not, to one of greater nuance such as "it is probable", "there is a possibility" or even "I don't know" – is required. Another pertinent case in point is the recent Gemini model debacle, i.e. when it was launched it was generating images of people which were adjusted to be representative and inclusive, but consequently ended up being context inaccurate. For example, when asked to generate images of Third Reich officers from WW2, it generated images of people of wide range of ethnicity, which would be factually incorrect [41]. This is a recent real life example of why PUQ in generative AI is important and how this must be matched to both context and demographic. The grand challenge here is that the allowable bounds of uncertainty for generative applications depend on the application at hand: whereas for creative applications like AI art or poetry a certain latitude in generative uncertainty (hallucinations) is permissible and even necessary, in other applications, such as clinical machine learning, where synthetic data is generated for data augmentation and class balancing, strict PUQ guarantees are needed at a individualised level.

4 Conclusion

In this Perspective, we have highlighted the need for PUQ in AI, and suggested eight grand challenges to developing and implementing PUQ in AI pipelines. Yet, PUQ will only have real impact if it is successfully communicated to the end user in an intuitive way. The benefits of communicating uncertainty are twofold: first, the decision-maker is able to understand the weight of their decision and make informed decisions that align with their values, expert judgment and risk tolerance. Second, the affected individual can better comprehend the level of confidence associated with the AI-driven recommendation, allowing them to make an informed decision to trust, or not, in the decision-making process [79]. Decisions informed by AI must present with clear and accessible representations of uncertainty [44]. User interfaces built on top of AI algorithms should consider visual aids that convey to the lay user the distribution of potential outcomes and the certainty, or lack thereof, underlying an individual prediction.

For example, in healthcare, AI algorithms may be deployed in many settings: to make personalized treatment recommendations [27], assist with emergency room triaging, or predict patient prognosis. PUQ can help clinicians grasp not only the AI prediction itself but also the magnitude of uncertainty associated with it, in order to weigh recommendations appropriately against their clinical judgment. Patients, increasingly interested in being informed about their care and medical journey, can also be provided with this information. As with all aspects of a clinical interaction, PUQ can enable clinicians to consider the risks, benefits, and the most critical information to convey, and tailor this information accordingly, ensuring that patients are adequately informed about the AI-driven aspects of their care. More generally, meeting the challenges we have presented could help design AI systems that accord with emerging AI safety legislation, such as the EU AI Act [30], and enable users to decide when and when not to adopt AI recommendations, and thereby make better decisions.

Declaration

The authors declare no conflict of interest. This work is supported by the Turing-Roche Strategic Partnership.

Author Contributions Statement

T Chakraborti led the work as both first author and coresponding author. T Chakraborti, CRS Banerji, C Harbron and B.D. MacArthur contributed to the conceptualisation and overall writing and editing of the paper. All other authors cotributed by writing parts and subsections of the paper.

Figure 1. Conformal prediction (CP) for personalised uncertainty quantification (PUQ). CP generates a prediction set that guarantees provable coverage of the ground truth for each individual prediction with a user-specified probability (subject to some technical caveats, described in the main text).

Figure 2. Grand Challenges in Personalised Uncertainty Quantification (PUQ) in a range of AI domains from classical predictive tasks to emerging applications

References

1. Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U Rajendra Acharya, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information fusion*, 76:243–297, 2021.

2. Ali Akbari and Roozbeh Jafari. Personalizing activity recognition models through quantifying different types of uncertainty using wearable sensors. *IEEE Transactions on Biomedical Engineering*, 67(9):2530–2541, 2020.

3. Ben Allen. The promise of explainable ai in digital health for precision medicine: a systematic review. *Journal of Personalized Medicine*, 14(3):277, 2024.

4. Bandar Almaslukh. A reliable breast cancer diagnosis approach using an optimized deep learning and conformal prediction. *Biomedical Signal Processing and Control*, 98:106743, 2024.

5. Anastasios N. Angelopoulos, Rina Foygel Barber, and Stephen Bates. Online conformal prediction with decaying step sizes. *Proceedings of Machine Learning research (PMLR)*, 235:1616–1630, 2024.

6. Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A unified review of theory and new challenges. *arXiv:2005.07972*, 2021.

7. Anastasios N Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

8. Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. *ProPublica*, 2016.

9. Gonul Ayci, Murat Sensoy, Arzucan Özgür, and Pinar Yolum. Uncertainty-aware personal assistant for making personalized privacy decisions. *ACM Transactions on Internet Technology*, 23(1):1–24, 2023.

10. Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.

11. Baptista. Zoo guide to network embedding, 2023.

12. R.F. Barber, E.J. Candes, A. Ramdas, and R.J. Tibshirani. Predictive inference with the jackknife+. *The Annals of Statistics*, 49(1):486–507, 2021.

13. Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023.

14. R.S.M. Barros and S.G.T.C. Santos. A large-scale comparison of concept drift detectors. *Information Sciences*, 451:348–370, 2018.

15. Manuela Battaglini and Steen Rasmussen. Transparency, automated decision-making processes and personal profiling. *Journal of Data Protection & Privacy*, 2(4):331–349, 2019.

16. F. Bayram, B.S. Ahmed, and A. Kassler. From concept drift to model degradation: An overview on performance-aware drift detectors. *Knowledge-Based Systems*, 245:108632, 2022.

17. Ginestra Bianconi. *Multilayer Networks*. Oxford University Press, July 2018.

18. Ginestra Bianconi. *Higher-Order Networks*. Cambridge University Press, November 2021.

19. Kevin M Boehm, Pegah Khosravi, Rami Vanguri, Jianjiong Gao, and Sohrab P Shah. Harnessing multimodal data integration to advance precision oncology. *Nature Reviews Cancer*, 22(2):114–126, 2022.

20. J. Buolamwini and T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, pages 77–91, 2018.

21. Stefano Canali, Viola Schiaffonati, and Andrea Aliverti. Challenges and recommendations for wearable devices in digital health: Data quality, interoperability, health equity, fairness. *PLOS Digital Health*, 1(10):e0000104, 2022.

22. Y. Cao, H. Peng, J. Wu, Y. Dou, J. Li, and P.S. Yu. Knowledge-preserving incremental social event detection via heterogeneous gnns. In *Proceedings of the Web Conference 2021*, pages 3383–3395, April 2021.

23. I. Y. Chen, F. D. Johansson, and D. Sontag. Why is my classifier discriminatory? In *Proceedings of the NIPS*, pages 3543–3554, 2018.

24. Jonas Christensen. Ai in financial services. In *Demystifying AI for the Enterprise*, pages 149–192. Productivity Press, 2021.

25. Kara Combs, Adam Moyer, and Trevor J Bihl. Uncertainty in visual generative ai. *Algorithms*, 17(4):136, 2024.

26. R.S.M. de Barros and S.G.T. de Carvalho Santos. An overview and comprehensive comparison of ensembles for concept drift. *Information Fusion*, 52:213–244, 2019.

27. Joseph DeFrank and Aline Luiz. Ai-based personalized treatment recommendation for cancer patients. *Journal of Carcinogenesis*, 21(2), 2022.

28. Armen Der Kiureghian and Ove Ditlevsen. Aleatory or epistemic? does it matter? *Structural safety*, 31(2):105–112, 2009.

29. Tiffany Ding, Anastasios Angelopoulos, Stephen Bates, Michael Jordan, and Ryan J Tibshirani. Class-conditional conformal prediction with many classes. *Advances in Neural Information Processing Systems*, 36, 2024.

30. European Union. Regulation (eu) 2024/1689. *Official Journal of the European Union*, 2024.

31. Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

32. S. Fatumo, T. Chikowore, A. Choudhury, M. Ayub, and K. Martin, A. R. & Kuchenbaecker. A roadmap to increase diversity in genomic studies. *Nature Medicine*, 27:24–29, 2021.

33. Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 10(2):455–482, 2021.

34. Joseph Futoma and colleagues. As good as it gets? a new approach to estimating possible prediction performance. *PLOS ONE*, 19(10):e0296904, 2024.

35. Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

36. J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia. A survey on concept drift adaptation. *ACM computing surveys (CSUR)*, 46(4):1–37, 2014.

37. S. Garg, S. Balakrishnan, Z.C. Lipton, B. Neyshabur, and H. Sedghi. Leveraging unlabeled data to predict out-of-distribution performance. In *NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications*, 2021.

38. Isaac Gibbs and Emmanuel Candes. Adaptive conformal inference under distribution shift. *Advances in Neural Information Processing Systems*, 34:1660–1672, 2021.

39. Isaac Gibbs and Emmanuel Candes. Conformal inference for online prediction with arbitrary distribution shifts. *Journal of Machine Learning Research (JMLR)*, 25.162:1–36, 2024.

40. Isaac Gibbs, John J Cherian, and Emmanuel J Candès. Conformal prediction with conditional guarantees. *arXiv preprint arXiv:2305.12616*, 2023.

41. Nico Grant. Google chatbot's a.i. images put people of color in nazi-era uniforms. *New York Times*, 2024.

42. Leying Guan. Localized conformal prediction: A generalized inference framework for conformal prediction. *Biometrika*, 110(1):33–50, 2023.

43. Gutknecht. Bits and pieces: understanding information decomposition from part-whole relationships and formal logic. *Journal*, 2021.

44. Vinyas Harish, Felipe Morgado, Ariel D Stern, and Sunit Das. Artificial intelligence and clinical decision making: the new nature of medical uncertainty. *Academic Medicine*, 96(1):31–36, 2021.

45. Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110:457–506, 2021.

46. H. Ibrahim, X. Liu, and A. D. Zariffa, N. & Morris. Health data poverty: an assailable barrier to equitable digital health care health data poverty: an assailable barrier to equitable digital health care. *The Lancet Digital Health*, 2:E666–E676, 2020.

47. Jackson. A complete characterisation of structured missingness, 2023.

48. Qianzhao Ji, Xiaoyu Jiang, Minxian Wang, Zijuan Xin, Weiqi Zhang, Jing Qu, and Guang-Hui Liu. Multimodal omics approaches to aging and age-related diseases. *Phenomics*, 4(1):56–71, 2024.

49. Neel Kanwal, Miguel López-Pérez, Umay Kiraz, Tahlita CM Zuiverloon, Rafael Molina, and Kjersti Engan. Are you sure it's an artifact? artifact detection and uncertainty quantification in histological images. *Computerized Medical Imaging and Graphics*, 112:102321, 2024.

50. Konrad J Karczewski and Michael P Snyder. Integrative omics for health and disease. *Nature Reviews Genetics*, 19(5):299–310, 2018.

51. M. V. Kiang, J. T. Chen, N. Krieger, C. O. Buckee, M. J. Alexander, J. T. Justin T. Baker, R. L. Buckner, G. Coombs III, K. W. Rich-Edwards, J. W.and Carlson, and J. Onnela. Sociodemographic characteristics of missing data in digital phenotyping. *Scientific Reports*, 11:14447, 2021.

52. Nikita Kotelevskii, Samuel Horváth, Karthik Nandakumar, Martin Takáč, and Maxim Panov. Dirichlet-based uncertainty quantification for personalized federated learning with improved posterior networks. *Intenrational Joint Conferences on Artificial Intelligence (IJCAI)*, 2023.

53. Ranganath Krishnan and Omesh Tickoo. Improving model calibration with accuracy versus uncertainty optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

54. S. Kumar and A. Srivastava. Bootstrap prediction intervals in non-parametric regression with applications to anomaly detection. In *Proc. 18th ACM SIGKDD Conf. Knowl. Discovery Data Mining*, 2012.

55. Jing Lei and Larry Wasserman. Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(1):71–96, 2014.

56. J.L. Lobo, J. Del Ser, A. Bifet, and N. Kasabov. Spiking neural networks and online learning: An overview and perspectives. *Neural Networks*, 121:88–100, 2020.

57. Tuve Löfström, Henrik Boström, Henrik Linusson, and Ulf Johansson. Bias reduction through conditional conformal prediction. *Intelligent Data Analysis*, 19(6):1355–1375, 2015.

58. Luca Longo, Mario Brcic, Federico Cabitza, Jaesik Choi, Roberto Confalonieri, Javier Del Ser, Riccardo Guidotti, Yoichi Hayashi, Francisco Herrera, Andreas Holzinger, et al. Explainable artificial intelligence (xai) 2.0: A manifesto of open challenges and interdisciplinary research directions. *Information Fusion*, 106:102301, 2024.

59. A. R. Martin, M. Kanai, Y. Kamatani, Y. Okada, and M. J. Neale, B. M. & Daly. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nature Genetics*, 51:1565–1576, 2019.

60. E. Matovu, B. Bucheton, J. Chisi, J. Enyaru, C. Hertz-Fowler, and M. & 38 others Koffi. Enabling the genomic revolution in africa. *Science*, 344:1260792, 2014.

61. Daily Milanés-Hermosilla, Rafael Trujillo Codorniú, René López-Baracaldo, Roberto Sagaró-Zamora, Denis Delisle-Rodriguez, John Jairo Villarejo-Mayor, and José Ricardo Núñez-Álvarez. Monte carlo dropout for uncertainty estimation and motor imagery classification. *Sensors*, 21(21):7241, 2021.

62. R. Mitra, S. F. McGough, T. Chakraborti, C. Holmes, R. Copping, N. Hagenbuch, S. Biedermann, J. Noonan, B. Lehmann, A. Shenvi, X. V. Doan, D. Leslie, G. Bianconi, R. Sanchez-Garcia, A. Davies, M. Mackintosh, E.-R. Andrinopoulou, A. Basiri, and B. D. Harbron, C. & MacArthur. Learning from data with structured missingness. *Nature Machine Intelligence*, 4:230, 2022.

63. C. Mougan and D.S. Nielsen. Monitoring model deterioration with explainable uncertainty estimation via non-parametric bootstrap. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37:12, pages 15037–15045, June 2023.

64. Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J.V. Dillon, B. Lakshminarayanan, and J. Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *stat*, 1050:17, 2019.

65. Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *Machine learning: ECML 2002: 13th European conference on machine learning Helsinki, Finland, August 19–23, 2002 proceedings 13*, pages 345–356. Springer, 2002.

66. Aleksandr Podkopaev and Aaditya Ramdas. Distribution-free uncertainty quantification for classification under label shift. In *Uncertainty in artificial intelligence*, pages 844–853. PMLR, 2021.

67. Wei Qian, Chenxu Zhao, Yangyi Li, Fenglong Ma, Chao Zhang, and Mengdi Huai. Towards modeling uncertainties of self-explaining neural networks via conformal prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 14651–14659, 2024.

68. Joaquin Quinonero-Candela et al. *Dataset Shift in Machine Learning*. The MIT Press, 2009.

69. Anil Rahate, Rahee Walambe, Sheela Ramanna, and Ketan Kotecha. Multimodal co-learning: Challenges, applications with datasets, recent advances and future directions. *Information Fusion*, 81:203–239, 2022.

70. Wendy A Rogers and Mary J Walker. Fragility, uncertainty, and healthcare. *Theoretical medicine and bioethics*, 37:71–83, 2016.

71. Yaniv Romano, Rina Foygel Barber, Chiara Sabatti, and Emmanuel Candès. With malice toward none: Assessing uncertainty via equalized coverage. *Harvard Data Science Review*, 2(2):4, 2020.

72. Yaniv Romano, Evan Patterson, and Emmanuel Candes. Conformalized quantile regression. *Advances in neural information processing systems*, 32, 2019.

73. Yaniv Romano, Matteo Sesia, and Emmanuel Candes. Classification with valid and adaptive coverage. *Advances in Neural Information Processing Systems*, 33:3581–3591, 2020.

74. Catherine Régis, Jean-Louis Denis, Maria Luciana Axente, and Atsuo Kishimoto. *Human-Centered AI: A Multidisciplinary Perspective for Policy-Makers, Auditors, and Users*. Chapman & Hall, London, 2024. ISBN 9780367359142.

75. Mauricio Sadinle, Jing Lei, and Larry Wasserman. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525):223–234, 2019.

76. R. Sebastiao and J. Gama. A study on change detection methods. In *Progress in artificial intelligence, 14th Portuguese conference on artificial intelligence, EPIA*, pages 12–15, October 2009.

77. R. Sebastiao and J. Gama. On evaluating stream learning algorithms. *Machine learning*, 90:317–346, 2013.

78. L. Seyyed-Kalantari, H. Zhang, M. B. A. McDermott, and M. Chen, I. Y. & Ghassemi. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature Medicine*, 27:111–120, 2021.

79. Harishankar Vasudevanallur Subramanian, Casey Canfield, Daniel B Shank, and Matthew Kinnison. Combining uncertainty information with ai recommendations supports calibration with domain knowledge. *Journal of Risk Research*, 26(10):1137–1152, 2023.

80. V. Suriyakumar and A. Narayanan. Title. In *Proceedings of the International Conference on Machine Learning*, volume 202, pages 7395–7405, 2023.

81. Arthur Thuy and Dries F Benoit. Explainability through uncertainty: Trustworthy decision-making with neural networks. *European Journal of Operational Research*, 317(2):330–340, 2024.

82. Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal prediction under covariate shift. *Advances in neural information processing systems*, 32, 2019.

83. Erico Tjoa and Cuntai Guan. A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE transactions on neural networks and learning systems*, 32(11):4793–4813, 2020.

84. T. Tsiampalis and D. B. Panagiotakos. Missing-data analysis: socio- demographic, clinical and lifestyle determinants of low response rate on self- reported psychological and nutrition related multi-item instruments in the context of the attica epidemiological study. *BMC Medical Research Methodology*, 20:262, 2020.

85. Victoria Volodina and Peter Challenor. The importance of uncertainty quantification in model reproducibility. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 379(2197), 2021.

86. Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*, volume 29. Springer, 2005.

87. A. D. Vyas and D. S. Eisenstein, L. G. & Jones. Hidden in plain sight — reconsidering the use of race correction in clinical algorithms. *The New England Journal of Medicine*, 382:2465–2474, 2020.

88. Bruno A Walther and Joslin L Moore. The concepts of bias, precision and accuracy, and their use in testing the performance of species richness estimators, with a literature review of estimator performance. *Ecography*, 28(6):815–829, 2005.

89. X. Wang and S. Kadıoğlu. Modeling uncertainty to improve personalized recommendations via bayesian deep learning. *International Journal of Data Science and Analytics*, 16:191–201, 2023.

90. D Ye, L Veen, A Nikishova, J Lakhlili, W Edeling, OO Luk, VV Krzhizhanovskaya, and AG Hoekstra. Uncertainty quantification patterns for multiscale models. *Philosophical Transactions of the Royal Society A*, 379(2197):20200072, 2021.

91. Margaux Zaffran, Olivier Féron, Yannig Goude, Julie Josse, and Aymeric Dieuleveut. Adaptive conformal predictions for time series. In *International Conference on Machine Learning*, pages 25834–25866. PMLR, 2022.

92. Eric Zhou and Dokyun Lee. Generative artificial intelligence, human creativity, and art. *PNAS nexus*, 3(3):pgae052, 2024.