Standardizing the Probabilistic Sources of Uncertainty for the sake of Safety Deep Learning

Axel Brando[1,*,†], Isabel Serra[1,†], Enrico Mezzetti[1,†], Francisco J. Cazorla[1,†] and Jaume Abella[1,†]

[1]Barcelona Supercomputing Center - Centre Nacional de Supercomputació (BSC-CNS), Plaça d'Eusebi Güell, 1-3, 08034 Barcelona, Spain

Abstract

Nowadays, critical functionalities are increasingly tackled by autonomous decision-making systems, which depend on Artificial Intelligence (e.g. Deep Learning) models. Still, most of these models are designed to maximize the generic performance rather than preventing potential irreversible errors. While robustness and reliability techniques have been developed, in the recent years, to fill this gap, the sources of uncertainty in those decision models are still ambiguous. With a view to standardizing the uncertainty sources, in this paper we present a formal methodology to disentangle those sources from a probabilistic viewpoint for any (regression and classification) supervised learning model. Once we associate a formula to each uncertainty type, we expose the terminology disagreement in the literature and we propose one that is aligned with other previous works. Finally, based on the proposed formulation, we present an integrated visualization method to represent all the uncertainty sources in a single figure to, ultimately, assisting the design of uncertainty-tailored actions.

Keywords

Safety, Uncertainty Sources, Supervised learning, Deep Learning, Aleatoric, Epistemic, Domain, Robustness, Reliability

1. Introduction

Artificial Intelligence (AI)-based Autonomous systems are the cornerstone of a revolution that involves a wide-range of fields including health [1, 2], self-driving cars [3, 4, 5], financial decisions [6, 7], industrial processes [8] and even AI protein folding [9, 10]. In this context, Deep Learning (DL) constitutes one of the main AI-based models given their strong capabilities of learning complex functions by maximizing a certain accuracy metric. Consequently, in recent years, a strong DL research line has emerged with the aim of constantly improving the general accuracy of such systems for the sake of adopting them in more and more real-world problems. However, when wrong decisions imply significant costs, ethical issues emerge [11, 12, 13] and the priorities regarding only obtaining "generic good estimators" change to "avoiding critical errors".

When considering Functional Safety (FUSA), normally regulated through domain-specific standards, critical errors must be mitigated to decrease their probability of occurrence, and safety measures need to be deployed to manage errors so that remedial actions can be taken before those errors (e.g., due to incorrect decisions) can cause any harm. Such process is provably effective only if the sources of uncertainty of the decisions taken can be properly identified and quantified.

Importantly, the *reliability* concept that motivates this paper is different than the *interpretability* of the decisions made by the system or, otherwise, the *causal effect analysis* that can be essential to build a complete safe system. The fundamental goal here is to develop a formal probabilistic methodology to identify which are the main sources of uncertainty for any (regression or classification) supervised learning problem.

Afterwards, the present article exposes the current disagreement that exists regarding the notation and identification of those uncertainty sources in the literature (e.g. between the uncertainty quantification community and the machine learning community) and proposes to use the Domain, Epistemic and Aleatoric (DEA) uncertainty disentanglement, by following and combining other literature references [14, 15, 16, 17, 18, 19, 20, 21].

Finally, based on the DEA mathematical formulation, this work introduces an integrated methodology to visualize all these sources of uncertainty in a regression plot with a view to recognize which uncertainty types are more clearly influencing the tackled problem at hand and to design uncertainty-tailored measures.

SafeAI 2023: The AAAI's Workshop on Artificial Intelligence Safety, Feb 13–14, 2023, Washington, D.C., US

*Corresponding author.

†These authors contributed equally.

axel.brando@bsc.es (A. Brando); isabel.serra@bsc.es (I. Serra); enrico.mezzetti@bsc.es (E. Mezzetti); francisco.cazorla@bsc.es (F.J. Cazorla); jaume.abella@bsc.es (J. Abella)

https://axelbrando.github.io/ (A. Brando); https://www.bsc.es/serra-mochales-isabel (I. Serra); https://www.bsc.es/ (E. Mezzetti); https://people.ac.upc.edu/fcazorla/ (F.J. Cazorla); https://people.ac.upc.es/jabella/ (J. Abella)

0000-0001-8103-391X (A. Brando); 0000-0003-4328-347X (I. Serra); 0000-0002-1886-2931 (E. Mezzetti); 0000-0002-3344-376X (F.J. Cazorla); 0000-0001-7951-4028 (J. Abella)

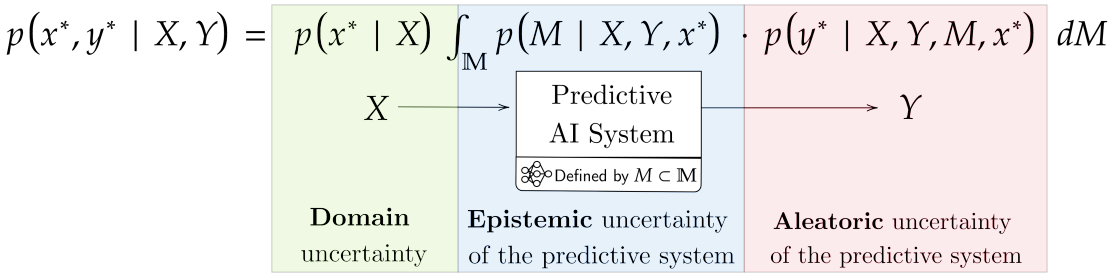

Figure 1: Proposed probabilistic disentanglement of the uncertainty sources for a supervised learning system (i.e. the predictive AI-based system). Note that only the last two types of uncertainty depend on the selected predictive system or model.

2. Formal Methodology

Generically for regression or classification problems, supervised learning models deal with two different kinds of random variables, (X, Y). The former one, X, which corresponds to the available information, and the latter one, Y, which is assumed to depend on X. The supervised goal is to find this probabilistic dependency.

In real-world problems, the joint distribution $p(X, Y)$ is a theoretical construction and we only have access to the data set, i.e. a certain population of instances, $D = \{(x_i, y_i)\}_{i=1}^{N}$, where each instance is assumed to be drawn i.i.d. from the aforementioned joint distribution. By using this existing data set, D, the supervised learning task corresponds to find a (probabilistic) function, $\phi_M : X \rightarrow Y$, by performing Empirical Risk Minimization (ERM) [22, 23, 24], which classically takes into account (1) the representativity - i.e., the assumed class of functions identified by its hyper-parameters, $M \sim \mathbb{M}$, (2) the optimization process, - i.e. which learning process is selected, and (3) the generalization capabilities, - i.e. evaluating the intra- and extrapolation ability of the learnt function, ϕ_M, over new scenarios.

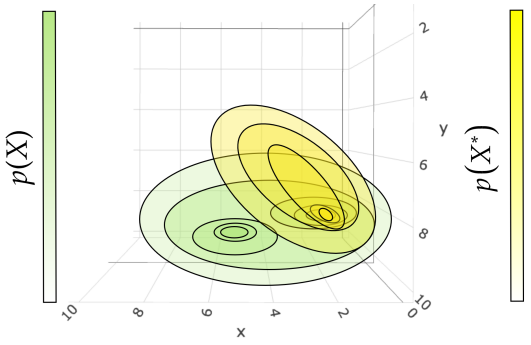

Figure 2: Level curves representation of the mismatch between a certain original distribution $p(X)$ and the new evaluated one, $p(X^*)$, which is our first uncertainty type tackled.

In particular, evaluating the generalization capabilities introduces the problem of detecting when the learnt function will be applied into a lower probability region with respect to $p(X)$. In other words, it is expected that a new evaluated point, x^*, could be sampled from a new random variable $x^* \sim X^*$ that hopefully will be similar to X. This difference is shown in Figure 2, which illustrates how it changes the "domain" or range of values affecting ϕ_M. In order to detect these changes, we should consider $p(x^* \mid X)$. At this point, Y is not considered.

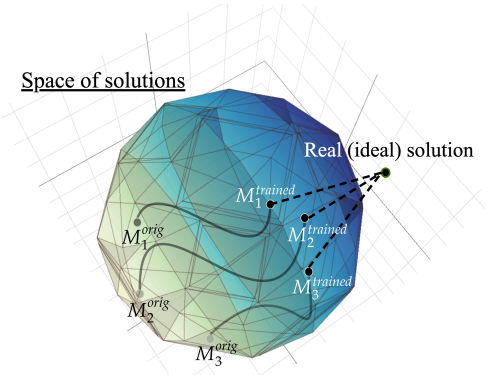

Figure 3: Representation of how the considered class of hyper-parameters, $\mathbb{M}$, constrains the models produced, identified by M and their proximity to the real (and ideal) solution.

Given that we are tackling a supervised problem, we are interested to evaluate the full pair of any new samples, (x^*, y^*), where the dependent random variable used in training, Y, should be also considered apart from X. Therefore, our goal will be to extend the previous conditional probability to compute $p(x^*, y^* \mid X, Y)$. Differently than in the previous $p(x^* \mid X)$ case, here in order to forecast the corresponding y^* value we require the predictive model, ϕ_M, which we assume it is identified by its hyper-parameters M. Following the representativity property nº (1) of the ERM, depending on the assumed class of functions (i.e. the space of solutions) character-

ized by $\mathbb{M}$, we can reach a closer or farther point to the hypothetical ideal solution, as it is shown in Figure 3. In that case, we are trying to estimate $p(M \mid X, Y, x^*)$ for all the possible $M \sim \mathbb{M}$, i.e. the set of models that maximize this conditional probability. Importantly, each of these models will produce a different response, - when it is evaluated as a function, - and the discrepancy between them is the uncertainty we are interested to capture here, which, in turn, is measuring the goodness of each model.

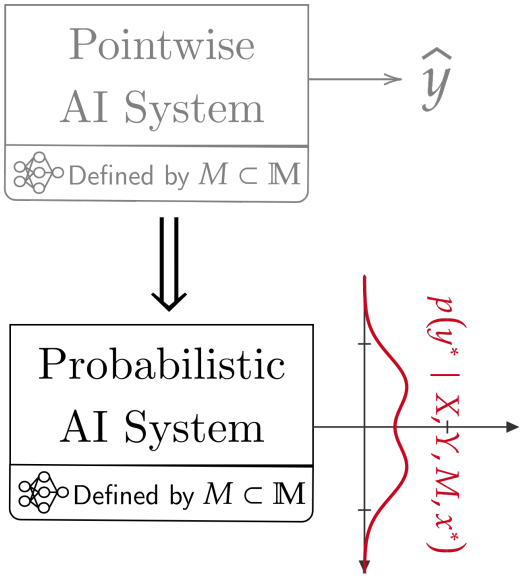

Figure 4: Required transformation of the supervised learning model from a point-wise forecaster to infer the conditional distribution, $p(y^* \mid x^*, X, Y, M)$.

Finally, non of the previous probability terms are considering the conditional variability of the dependent variable, y^*, given the data set, (X, Y), the desired model, ϕ_M, and the input information to forecast, x^*, i.e. $p(y^* \mid X, Y, M, x^*)$. This conditional probability includes effects such as occlusion or lack of input variables to unambiguously forecasting the y^* value. In order to tackle it, we require that our model does not outputs a deterministic pointwise prediction but it should characterizes the aforementioned conditional distribution, as shown in Figure 4.

On the whole, the three described probabilities are the terms of the chain rule of the following joint distribution,

$$p(x^*, y^* \mid X, Y) = p(x^* \mid X) \cdot \\ \cdot \int_{\mathbb{M}} p(M \mid x^*, X, Y) \cdot p(y^* \mid x^*, X, Y, M) dM \quad (1)$$

which is graphically represented over the predictive AI system in Figure 1. These three terms of Eq. (1) are the types of uncertainty we will follow hereafter. Once we have associated a formula to each one, we will now enter into the terminological discussion around their names.

3. Uncertainty Types Taxonomy

Understanding the sources of the uncertainty is not strictly a data science or machine learning concern. In fact, their characterization is the main goals of the uncertainty quantification science [25] but also a major issue in many fields in natural sciences, engineering and even constitute a philosophical debate [26, 27].

One of the more extended taxonomy is the one that disentangles the types of uncertainty depending on their are reducible or not [28]. Theoretically, we can reduce uncertainty by *improving* our observational data and experimental techniques (e.g. we can increase the measurement precision, we can collect more high-quality data or we can consider a more proper family of models). However, realistically, there are problems where actual limitations or assumptions (e.g. time or resources constrains) prevent us to perform these improvements. In these scenarios appear irreducible uncertainties. In the literature [29], when the uncertainty is reducible it is called "Epistemic uncertainty" and when is irreducible it is called "Aleatoric uncertainty".

Compared with Figure 1, this reducible-irreducible dichotomy do not explicitly refers to the use of a predictive model. In particular, we can observe that the last term of Eq. (1), $p(y^* \mid x^*, X, Y, M)$, is clearly Aleatoric - due to it is focused on modelling the irreducible conditional variability of the response variable -. Similarly, the inherent noise in the input variables, X, - e.g. due to an precision measuring error, - it is also irreducible in realistic problems, therefore, it is Aleatoric in these terminology[1]. On the other hand, the reducible uncertainty can be identified into the data itself, - which consist to the $p(x^* \mid X)$ term -, or also when the model is considered, which includes the marginalization over M and the $p(M \mid x^*, X, Y)$ term.

On the whole, there is an Aleatoric and Epistemic uncertainty for the data part and an Aleatoric and Epistemic uncertainty for the model part of any supervised learning scheme.

The fact that different meanings of uncertainty use the same terminology exacerbates the debate on the appropriate use or not of the terms "Epistemic" and "Aleatoric" for this disentanglement. For instance, in the case of

[1] Importantly, in the uncertainty disentanglement presented in this paper, we are considering this kind of irreducible uncertainty negligible. Therefore, it is not present in Eq. (1)

Epistemic uncertainties, $p(M \mid x^*, X, Y)$ [2], is also identified with terms like "model", "procedural", "parameter", "systematic" uncertainty, which is different than the $p(x^* \mid X)$ that can be named as "distribution shift", "data uncertainty", "outlier detection" or "out-of-distribution detection". Similarly, the Aleatoric $p(y^* \mid x^*, X, Y, M)$ is also known as "occlusion", " statistical", "random" or even "lack-of-rows" (when a table-viewpoint is applied), which is different than the measurement-noise or residual variability that is non-conditional.

Given that we are assuming this (non-conditional) inherent noise in the input variable X as negligible, in consonance with other approaches [30, 31, 32, 33, 19], in this paper we decided to use this irreducible and non-irreducible terms only for the model part. Therefore, in our terms, *Epistemic uncertainty* will correspond to the integral and $p(M \mid x^*, X, Y)$ of Eq. (1) and the *Aleatoric uncertainty* will correspond to the $p(y^* \mid x^*, X, Y, M)$ term of Eq. (1), as shown in Figure 1.

Finally, we require a name for the $p(x^* \mid X)$ term of Eq. (1). Based on other works [18], we name it as "domain uncertainty". Therefore, the presented framework of this paper (see Figure 1) will consider the Domain, the Epistemic and the Aleatoric (DEA) uncertainty sources disentanglement and, since they come from an expression that integrates them all, this expression directly indicates us how the different types of uncertainty can be combined regardless of the dimensionality of X or Y and obtaining a single integrated uncertainty term, $p(x^*, y^* \mid X, Y)$.

4. DEA as a Single Integrated Term

Importantly, the previously exposed DEA disentanglement does not assume any specific supervised task such as regression or classification. Furthermore, the dimensionality of X and Y is not specified. The only crucial point is to compute the Eq 1 integrated term $p(x^*, y^* \mid X, Y)$ and use this information to build reliable DL-based models (e.g. conditioning the human intervention depending on that uncertainty information).

Moreover, the disentanglement allows the possibility to characterize the tackled problem, by analysing which of the DEA uncertainties dominate over the others, as proposed in [21]. Generically speaking, a problem where the detected Aleatoric uncertainty, $p(y^* \mid x^*, X, Y, M)$, prevails over the other ones indicates that the input information X omits relevant variables for the forecasting task, consequently, a point-wise approximation would neglect critical information. In contrast, if the predicted uncertainty is mostly the Epistemic one, then it can indicates that the considered family of models could be too restrictive or there exist a notable model bias. Finally, if the Domain uncertainty monopolize the integrated value, then we would apparently be facing frequent anomalies in the test phase, which may indicate us that the test distribution is differently than the training one, e.g. because suddenly there was a distribution change between the original input information, X, and the new one to be evaluated, X^*.

[2] Here the marginalization of M is implicitly considered.

5. Visualizing the DEA integration

Until now, the DEA disentanglement introduces a procedure to split the uncertainty sources depending on the associated probability it has, which are obtained using the chain rule. Based on this and following [21], in the next subsections we will describe how to visualize each uncertainty type and how to merge them all in a single visualization plot for a regression problem with a single dimension for X and Y to be visualized in a 2D plot[3].

5.1. The Domain uncertainty

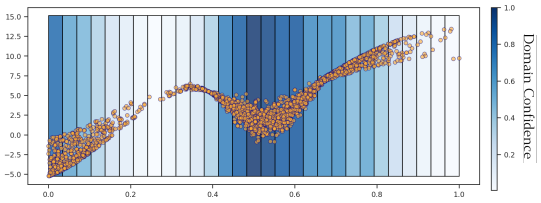

Figure 5: The Domain uncertainty is captured by modelling $p(X)$. This value is shown in the background in blue.

First of all, it is important to highlight that the Domain uncertainty does not consider the variable to be predicted, Y, nor the model to perform such prediction, characterized by M. Therefore, we should be careful when we represent this uncertainty in a standard regression plot where the horizontal axis is some input variable and the vertical axis is the predicted values, due to $p(x^* \mid X)$ not depend on Y. Consequently, one way to represent this uncertainty can be using the background colour as represented in Figure 5. In that case, each horizontal value has a different background colour where purpler zones correspond to high confidence X values while zones with lower X confidence will be bluish.

Analysing Figure 5 we can see that zones where there are less data points have a whitish colour (e.g. between 0.2 and 0.4 or between 0.8 and 1). Furthermore, we should highlight that the $p(x^* \mid X)$ value is independent of the conditional variability $p(y^* \mid x^*, X, Y)$, as we can observe, for instance, between the 0.4 and the 0.6 points.

[3] As a position paper, the implementation details are omitted here for an extended version of the article to focus now in the combination procedure of the DEA uncertainties. Importantly, this DEA combination procedure does not depend on the X or Y dimensionality.

5.2. The Epistemic uncertainty

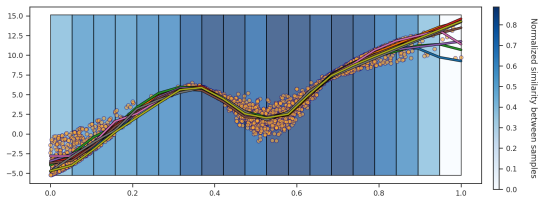

Figure 6: Each NN component of the ensemble is approximating the conditional median. The discrepancy on those components encodes the Epistemic uncertainty. In blue, the normalized standard deviation is shown in the background.

Epistemic uncertainty, $p(M \mid x^*, X, Y)$, which corresponds to the uncertainty related to selecting a certain family of models $\mathbb{M}$ (see Figure 3) is a similar case than the Domain uncertainty: The new response random variable y^* is not involved in this uncertainty. However, here the prediction of each model, - characterized by M, - is usually approximating some statistic of Y given X[4]. This last detail produces several ways of visualizing this uncertainty depending on what is approximating each model but we should be careful to distinguish between the Epistemic and Aleatoric approximation part. To do it, here we will consider only point-wise approximator models, e.g. a model that is predicting the conditional median.

At the end, considering a certain family $\mathbb{M}$ of such models is to consider an ensemble with a finite - or not - number of components or models. Their discrepancy refers to the Epistemic uncertainty we are capturing. Therefore, one way to visualize its discrepancy is to plot each prediction separately, as it is shown in Figure 6.

Comparing Figure 6 with Figure 5, we can see clearly the difference between capturing the Domain or Epistemic uncertainty: For instance, in the horizontal interval from 0.2 to 0.4, the behaviour of both uncertainties are completely different. This is because the density of $p(X)$ is small in such interval but, differently, the approximated conditional medians of the ensemble are producing a similar forecast given the previous and posterior shape of the data is clearly defined (and the consequence behaviour that use to perform NN models). This fact could tend to change when X is high dimensional but, if the ensemble is naively approximated, we do not have any guarantee that the discrepancy will be always higher in zones where $p(X)$ is lower using such NN models.

Importantly, similarly to the Domain case, it is worth to highlight that high conditional variability zones between the response variable and the input one, such as in the horizontal interval from 0.4 to 0.6, does not imply to having an Epistemic discrepancy if the approximated statistic is clearly defined. Therefore, we should need to model Aleatoric uncertainty to detect this extra new source of uncertainty.

[4]The standard gold approach in regression problems is to approximate the conditional mean, - or median -, which comes from minimizing the mean square error [34], - or mean absolute error, respectively [35]-.

5.3. The Aleatoric uncertainty

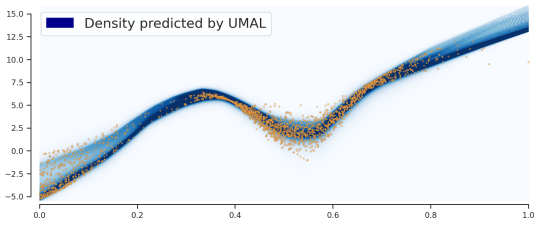

Figure 7: Conditional distribution approximated using the UMAL model [32]. Aleatoric uncertainty is captured as the approximated likelihood.

Our Aleatoric uncertainty is focused on modelling the conditional variability of the response variable $p(y^* \mid x^*, X, Y, M)$ of Eq. (1).

Unlike previous uncertainty types, visualizing Aleatoric uncertainty has a direct impact on the response variable to be modelled, therefore, this uncertainty can be represented without the vertical bars used in Figure 5 and 6 because now it depends on the vertical axis values. Additionally, this uncertainty is irreducible, therefore, our goal will be to show the distributional shape to design shape-tailored techniques. This is why it is important to avoid strong assumptions regarding to the conditional distribution, such as symmetry or unimodality, if we do not have clear evidences that they cannot harm the forecasting or visualization procedure, as it is deeply discussed in [21].

To provide a richer estimation of the likelihood beyond the standard aleatoric conditional symmetric and unimodal approaches [30], in Figure 7 we can see a UMAL forecast[5] [32] of the previously presented data set, where blueish areas are the ones that has higher likelihood.

Overall, we can observe that lower likelihood points are those where the conditional variability is higher. Therefore, between $[0., 0.2]$, $[0.4, 0.6]$ and $[0.8, 1]$. This behaviour contrast with the presented in previous Figures 5 and 6 as we will discuss in the next subsection when an integrated approach will be designed.

Importantly, isolated aletoric uncertainty fixes a certain model parametrized by M. This can be seen as one

[5]The UMAL model learns a conditional mixture of an infinite number of Asymmetric Laplacians using a neural network. Therefore, it can learn multi-modalities and asymmetries if they appear.

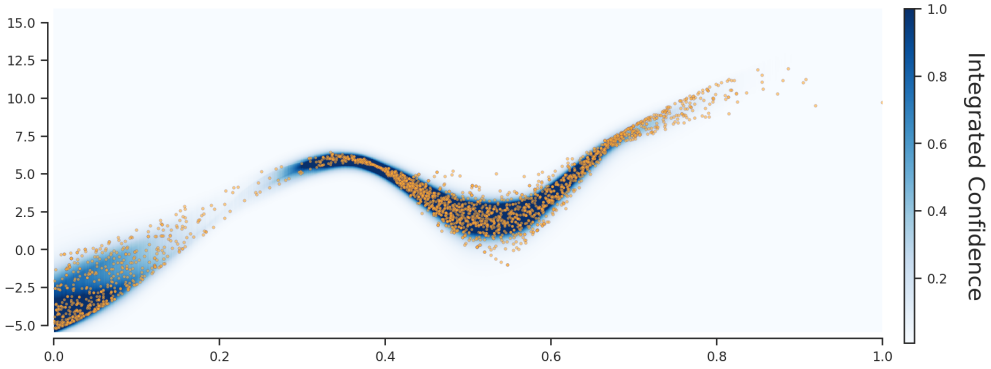

Figure 8: Continuous Integrated uncertainty visualization based on Figures 5, 3 and 7. All uncertainties are considered as probabilities.

of the components of the ensemble in the previous Epistemic subsection and, based on this idea, we can build an integration procedure to visualize all the presented uncertainty sources as follows.

5.4. The integrated visualization

Proposing an integrated procedure to represent all the uncertainty types is useful to synthesize all this complex information in a single plot. Based on the previously introduced visualization types, we can represent the presented Domain, Epistemic and Aleatoric uncertainties in an integrated visualization using Eq. 1, as shown in Figure 8. This representation displays the confidence in all the uncertainty sources using the $p(x^*, y^* \mid X, Y)$ estimated information, which includes modelization of the outlier detection, model uncertainty and conditional irreducible uncertainty.

6. Conclusion

In recent times, technological advances have led to increase the use of Artificial Intelligence (AI) systems for critical autonomous decision-making. This emphasizes the importance of developing robust and reliable methods, where determining the sources of the uncertainty constitutes an essential pillar.

Still, in the literature, there is not a clear standard to identify the sources of the uncertainty. To support this objective, in this paper we have presented the Domain, Epistemic and Aleatoric (DEA) disentanglement; a formal methodology to divide the uncertainty sources from a probabilistic viewpoint for any (regression or classification) supervised learning model. Furthermore, we presented a comparison of the DEA disentanglement to other literature nomenclatures and approaches.

Finally, the proposed unified approach launches the possibility to combine all the literature uncertainty modelling techniques. Furthermore, it provides an integrated procedure to visualize them together for the sake of recognizing which is the effect of each uncertainty type.

Overall, we hope that the presented framework can help to build a backbone where to connect future research in designing autonomous AI-based systems that requires Safety certifications and to tackle other needs beyond.

Acknowledgments

The research leading to these results has received funding from the Horizon Europe Programme under the SAFEXPLAIN Project (www.safexplain.eu), grant agreement num. 101069595 and the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 772773). Additionally, this work has been partially supported by Grant PID2019-107255GB-C21 funded by MCIN/AEI/ 10.13039/501100011033.

References

[1] R. C. Deo, Machine learning in medicine, Circulation 132 (2015) 1920–1930.

[2] A. Rajkomar, J. Dean, I. Kohane, Machine learning in medicine, New England Journal of Medicine 380 (2019) 1347–1358.

[3] J. Stilgoe, Machine learning, social learning and the governance of self-driving cars, Social studies of science 48 (2018) 25–56.

[4] R. Michelmore, M. Kwiatkowska, Y. Gal, Evaluating uncertainty quantification in end-to-end autonomous driving control, arXiv preprint arXiv:1811.06817 (2018).

[5] H. Tunga, R. Saha, S. Kar, A method of fully autonomous driving in self-driving cars based on machine learning and deep learning, Intelligent Multimodal Data Processing (2021) 131–156.
[6] R. Culkin, S. R. Das, Machine learning in finance: the case of deep learning for option pricing, Journal of Investment Management 15 (2017) 92–100.
[7] J. De Spiegeleer, D. B. Madan, S. Reyners, W. Schoutens, Machine learning for quantitative finance: fast derivative pricing, hedging and fitting, Quantitative Finance 18 (2018) 1635–1643.
[8] A. Diez-Olivan, J. Del Ser, D. Galar, B. Sierra, Data fusion and machine learning for industrial prognosis: Trends and perspectives towards industry 4.0, Information Fusion 50 (2019) 92–111.
[9] F. Noé, G. De Fabritiis, C. Clementi, Machine learning for protein folding and dynamics, Current opinion in structural biology 60 (2020) 77–84.
[10] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, et al., Highly accurate protein structure prediction with alphafold, Nature 596 (2021) 583–589.
[11] A. Hevelke, J. Nida-Rümelin, Responsibility for crashes of autonomous vehicles: an ethical analysis, Science and engineering ethics 21 (2015) 619–630.
[12] P. Lin, Why ethics matters for autonomous cars, in: Autonomous driving, Springer, Berlin, Heidelberg, 2016, pp. 69–85.
[13] A. Maxmen, Self-driving car dilemmas reveal that moral choices are not universal, Nature 562 (2018) 469–469.
[14] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U. R. Acharya, et al., A review of uncertainty quantification in deep learning: Techniques, applications and challenges, Information Fusion 76 (2021) 243–297.
[15] W. J. Maddox, P. Izmailov, T. Garipov, D. P. Vetrov, A. G. Wilson, A simple baseline for bayesian uncertainty in deep learning, Advances in Neural Information Processing Systems 32 (2019).
[16] A. Loquercio, M. Segu, D. Scaramuzza, A general framework for uncertainty estimation in deep learning, IEEE Robotics and Automation Letters 5 (2020) 3153–3160.
[17] Y. Gal, Z. Ghahramani, Dropout as a bayesian approximation: Representing model uncertainty in deep learning, in: international conference on machine learning, PMLR, 2016, pp. 1050–1059.
[18] A. Ashukha, A. Lyzhov, D. Molchanov, D. Vetrov, Pitfalls of in-domain uncertainty estimation and ensembling in deep learning, arXiv preprint arXiv:2002.06470 (2020).
[19] E. Hüllermeier, W. Waegeman, Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods, Machine Learning 110 (2021) 457–506.
[20] H. Lin, Bayesian epistemology (2022).
[21] A. Brando, Aleatoric uncertainty modelling in regression problems using deep learning, Universitat de Barcelona (2022).
[22] V. Vapnik, Principles of risk minimization for learning theory, Advances in neural information processing systems 4 (1991).
[23] J. Berner, P. Grohs, A. Jentzen, Analysis of the generalization error: Empirical risk minimization over deep artificial neural networks overcomes the curse of dimensionality in the numerical approximation of black–scholes partial differential equations, SIAM Journal on Mathematics of Data Science 2 (2020) 631–657.
[24] H. Zhang, M. Cisse, Y. N. Dauphin, D. Lopez-Paz, mixup: Beyond empirical risk minimization, in: International Conference on Learning Representations, 2018.
[25] T. J. Sullivan, Introduction to uncertainty quantification, volume 63, Springer, 2015.
[26] I. Hacking, et al., The emergence of probability: A philosophical study of early ideas about probability, induction and statistical inference, Cambridge University Press, 2006.
[27] J. Von Plato, Creating modern probability: Its mathematics, physics and philosophy in historical perspective, Cambridge University Press, 1994.
[28] A. Der Kiureghian, O. Ditlevsen, Aleatory or epistemic? does it matter?, Structural safety 31 (2009) 105–112.
[29] H. G. Matthies, Quantifying uncertainty: Modern computational representation of probability and applications, in: Extreme man-made and natural hazards in dynamics of structures, Springer, 2007, pp. 105–135.
[30] A. Kendall, Y. Gal, What uncertainties do we need in bayesian deep learning for computer vision?, Advances in neural information processing systems 30 (2017).
[31] N. Tagasovska, D. Lopez-Paz, Single-model uncertainties for deep learning, Advances in Neural Information Processing Systems 32 (2019).
[32] A. Brando, J. A. Rodriguez, J. Vitria, A. Rubio Muñoz, Modelling heterogeneous distributions with an uncountable mixture of asymmetric laplacians, Advances in neural information processing systems 32 (2019).
[33] G. Li, L. Yang, C.-G. Lee, X. Wang, M. Rong, A bayesian deep learning rul framework integrating epistemic and aleatoric uncertainties, IEEE Transactions on Industrial Electronics 68 (2020) 8829–8841.
[34] R. Moore, J. DeNero, L1 and l2 regularization for

multiclass hinge loss models, in: Symposium on machine learning in speech and language processing, 2011.

[35] M. Schmidt, G. Fung, R. Rosales, Fast optimization methods for l1 regularization: A comparative study and two new approaches, in: European Conference on Machine Learning, Springer, 2007, pp. 286–297.